UNIVERSITY OF ALBERTA

HYBRID CLASSIFICATION SYSTEMS WITH EVIDENCE THEORY

By

Trang Huyen Le

A Thesis Submitted to the Faculty of Graduate Studies and Research
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

Department of Electrical & Computer Engineering

Edmonton, Alberta, Canada

Fall 2007

Library and Archives Canada
Published Heritage Branch
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada
Direction du Patrimoine de l'édition
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-33292-4
Our file Notre référence
ISBN: 978-0-494-33292-4

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

ABSTRACT

This thesis proposes a novel adaptive classifier that is suited to deal with the uncertainty in the process of assigning labels for unknown patterns. The novel classifier is built based on two conceptual levels which are global and local levels. The global level provides insights about the given data set from the global point of view. In addition, the local level supplies extra perceptions about the same data set from the local point of view. The significant improvement of the proposed classifier is that knowledge obtained from the global and local level is represented in evidence–theoretical framework and combined by Dempster–Shafer combination rules, respectively. As the results, the labeling tasks for unknown patterns rely on the level of confidence which comprehends as beliefs support for a specific category related to a particular pattern or as ignorance presented for the lacks of knowledge when deciding the precise classification for a specific sample. Applicability of the model is validated with three different data sets.

ACKNOWLEDGMENT

I would like to express my sincere thanks to my co-supervisors, Drs. Marek Reformat and Witold Pedrycz, for their invaluable feedback, guidance and supports, as well as for sharing with me the knowledge and experiences. I really appreciate their time and painstaking efforts for providing me detail markups for thesis outline and numerous key ideas which use throughout this research.

My husband, Cuong Ly, my parents, Hai Mau Le and Yen Kim Tran, my sister Ngoc Bich Le for their constant love and support. They encourage me and could be counted on to uplift my spirits and my motivation whenever I needed it.

Finally, I would like to express my appreciation to the administrative and technical support staff for their valuable assistance, and for providing a conducive environment in which I could carry out my study.

CONTENTS

1. INTRODUCTION 1

1.1 OVERVIEW 1

1.2 MOTIVATION 2

1.3 THESIS CONTRIBUTION 4

1.4 THESIS OUTLINE 5

2. THEORETICAL BACKGROUND 7

2.1 FUZZY SETS 7

2.2 CLUSTERING 14

2.2.1 Basic concepts 14

2.2.2 Clustering Algorithms 19

2.2.2.1 Hard–Cmeans Clustering 19

2.2.2.2 Fuzzy–Cmeans Clustering 22

2.2.3 Clustering – based Classification 25

2.3 EVIDENCE THEORY 28

2.3.1 Basic Concepts 28

2.3.2 Dempster – Shafer Combination Rules 35

3. RELATED WORKS 41

4. APPLICATION OF EVIDENCE THEORY TO CLUSTERING-BASED CLASSIFICATION 46

4.1 OVERVIEW 46

4.2 LOCAL-BASED CLASSIFICATION 49
4.2.1 Concept 49
4.2.2 Calculation of Belief Masses 50
4.2.3 Local Beliefs 51
4.3 GLOBAL-BASED CLASSIFICATION 52
4.3.1 Concept 52
4.3.2 Calculation of Belief Masses 54
4.3.3 Global Beliefs 56
4.4 COMBINING LOCAL AND GLOBAL BELIEFS 60
5. EXPERIMENTAL RESULTS 62
5.1 Synthetic Data 64
5.1.1 Data Generation 64
5.1.2 Results 67
5.2 Synthetic Data 75
5.2.1 Data Generation 75
5.2.2 Results 76
5.3 Real World Data 81
5.3.1 Data Description 81
5.3.2 Results 82
6. CONCLUSIONS 85

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

With the advance in both hardware and software technologies, process of data generation and storage has become faster than ever [11]. For example, WalMart, a U.S retailer has biggest business database in the world with over 20 million transactions per day [3]. NASA Earth Observing System (EOS) of orbiting satellites and other space-borne instruments generate 50 gigabytes of remotely sensed image data per hour [28]. Obviously, such huge amount of data is beyond human ability to analyze manually. Thus, the seek for novel tools or techniques that can automatically analyze and extract useful information from databases becomes crucial. It leads to the birth of the new fields named Knowledge Discovery in Databases (KDD) and Machine Learning (ML) which attract great interests from researchers

in many different fields.

KDD process contains several steps from data processing to data evaluation in which data mining is consider as the most important step. The role of data mining step is to use a well–defined data mining and machine learning techniques such as classification, clustering, regression, etc to discover hidden information from data. Based on discovered knowledge, models or descriptors are built which serve for prediction or descriptions of future data sets.

Even though, data mining and machine learning techniques prevail in many applications such as data processing, decision making, control problems, it is easy to be recognized that most models have been designed to handle accurate data; whereas in the real life data that we encounter are often imprecise and uncertain [31]. Growing awareness of dealing with uncertainty could help the new model to avoid of making ill inferences of a given data set. Hence, it is very beneficial to find new techniques where uncertainty can be managed properly.

1.2 MOTIVATION

Classification in the context of machine learning is to find a function that maps a data item into one of several predefined classes [30]. The classification task is to analyze the training data to develop accurate descriptions or to build a model according to the relationships between classes and present features in the data set.

The future data then can be classified using the class descriptions or models which use to provide more knowledge about each class in the data set.

Clustering is to group a given data set into clusters based on similarities among features present in the data set [16], [26]. These clusters is considered as structures of the data set. According to these structures, hidden information or assumptions about the data set is revealed or made respectively, depending on application domains.

Since classification and clustering become two important techniques in data analysis, they still cope with some limitations of dealing with uncertainty. Even though, Bayesian inference has been proposed to overcome these drawbacks, there are still issues related to combining multiple information that comes from different sources and building a more effective classifier.

This research is motivated to build a novel classifier to refine the above issue. This novel classifier is constructed by taking advantage of some of the classification and clustering methods developed in the past and extend them with evidence theory. Moreover, the proposed classifier provides insights about the data on two conceptual levels. Based on the confidence degree obtained from the combination of two conceptual levels, we can gain more precise conclusions for classification with respect to questionable patterns.

1.3 THESIS CONTRIBUTION

In the proposed thesis, a new approach for constructing classification systems is proposed and investigated. The new method is based on a combination of two different classification techniques: k-Nearest Neighbors, and clustering based classification. The procedure that is applied to "combine" these two classifiers is based on evidence theory. To make this possible both techniques have been augmented.

The kNN approach is using beliefs that are calculated based on distances to neighbors. This approach represents a local view on classification – a decision is reached based on points that are close to a point that is being classified.

The cluster–based classification is also modified to incorporate beliefs. In this case the beliefs are evaluated using distances to the centroids of clusters. This approach represents a global view on classification – a decision is based on data patterns that are discovered by clustering. In order to eliminate a problem of identifying a suitable number of clusters, the clustering process is repeated multiple times. Every time a different number of clusters is used. Classification results obtained for each run are combined using elements of evidence theory.

The new approach that merges both techniques equipped with elements of evidence theory shows great advantages in data classification.

In the proposed thesis a number of data are analyzed using the proposed

technique. The principles of the technique have been explained in the case of two-dimentional synthetic data with three categories.

1.4 THESIS OUTLINE

The remainder of this thesis is organized as follows:

In Chapter 2, fundamental concepts about fuzzy sets, clustering, evidence theory and Dempster–Shafer combination rules, classification and clustering–based classification techniques are given, along with some algorithms developed for clustering methodology, several applications of clustering–based classification techniques in different domains.

In Chapter 3, we briefly survey related works regarding to the application of Dempster–Shafer combination rules on knowledge obtained from data mining techniques. Summary of their works, key ideas also experimental results are reported.

In Chapter 4, a novel model for classification is proposed. The process of achieving knowledge from global and local levels is described carefully in this part. Important figures are plotted to illustrate key ideas of how the model works and how the data set is treated inside the new classifier. Furthermore, equations to represent knowledge obtained from global and local levels in the evidence–theoretical framework are defined, as well as the process of combining them by applying Dempster–Shafer combination rules on appropriate steps.

In Chapter 5, applicability of the proposed classifier are implemented, tested and the results, both on synthetic databases and real databases are reported. Several comments related to the results obtained from the model are made.

The major work is summarized and the conclusions from the study are given. Some possible future works are also discussed.

CHAPTER 2

THEORETICAL BACKGROUND

2.1 FUZZY SETS

Since the appearance of mathematics becomes a major discipline in daily life; human being starts getting familiar with the concept of high standard precision. Roughly speaking, the high standard precision deals with the sharp, well–defined boundary or quantitative measure for a specific phenomenon. Later, the high standard precision is conceptualized to be crisp sets in which each object in a given collection (or set) has either complete belonging or complete not belonging scenarios. Although, crisp sets prevail in many areas of applications such as mathematics, chemistry and so on, it can easily be seen that it lacks the flexibility to handle imprecision and vagueness [32].

Real life usually encounters with uncertainty, incomplete information; especially in human semantics when making descriptions, judgments or decisions about

a specific problem. Notions such as "tall", "very fast" can not handle by crisp sets because the interpretation of these notions depends on particular contexts or people's knowledge. In addition, in recent decades, the advent of computer science endeavoring to build human-centre systems which can mimic the way of human thinking stimulates researchers to look for a novel, appropriate notion which can cope with uncertainty properly.

Fuzzy sets were initiated by Zadeh in 1965, a professor in computer science department at the University of California in Berkeley, as the solution to address all issues above [33]. Fuzzy sets are the extension of crisp sets by associating a degree of membership to each object in the set. The degree of membership for each object is in the range $[0, 1]$. The higher the membership value, the more typical the object belongs to the set [33]. With this characteristic, the fuzzy set boundary changes gradually not abruptly as in crisp sets. In other words, fuzzy set boundary is soft and extendable. Therefore, fuzzy sets are sufficiently flexible to deal with uncertainty or imprecise information.

In this section, very basic definitions of fuzzy sets and their operations are reviewed.

Definition

Since long time, a crisp set is a foundation for traditional mathematic. A crisp

set A in the universe X is defined by a characteristic function:

$$A : X \rightarrow \{0, 1\}$$

in which each element $x \in X$ is assigned a number either 0 or 1. The number 0 denotes for complete exclusion of the set A and the number 1 denotes for complete member of the set A [21]. In some other references [21], [34], the characteristic function is expressed as:

$$A(x) = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \notin A \end{cases}$$

This idea is extended to fuzzy sets. The main difference from a crisp set and a fuzzy set is that: in the case of a crisp set, there is no gradations of belonging; whereas, in the case of a fuzzy set, there is gradations of belonging. More specifically, with a crisp set, there is nothing in between belonging and not belonging; while on the contrary, with a fuzzy set, each member shows its degree of belonging to a set A by associating itself with a membership value in the interval $[0, 1]$.

More mathematically, a fuzzy set A of the universe X is characterized by a membership function [34]:

$$\mu_A : X \rightarrow [0, 1]$$

Each member in the universe X together with its membership value is called a fuzzy singleton. For instance, if $A = \{x\}$ and x is supported with a membership

value μ, A is represented as:

$$A = \mu/x$$

If the set A contains a finite numbers of member $\{y_1, y_2, \cdots, y_n\}$ with corresponding membership values $\{\mu_1, \mu_2, \cdots, \mu_n\}$ respectively, A is defined by a summation of all fuzzy singletons [34]:

$$A = \mu_1/y_1 + \mu_2/y_2 + \cdots + \mu_n/y_n$$

or

$$A = \sum_{i=1}^{n} \mu_i/y_i$$

where μ_i is the degree of membership for each member y_i.

Operations on Fuzzy sets

Since Union, Intersection and Complement operations are well–known in classical mathematics, those operations are still kept in fuzzy framework. Besides, the strength of fuzzy systems is the ability to capture or express various levels of linguistic intensities; therefore, some operations as Concentration, Dilation, Contrast Intensification, Fuzzification, Product are essential to support this characteristic [34].

Intersection Operation

The intersection of two sets A and B is denoted as $A \cap B$ and is defined by:

$$A \cap B \triangleq \int_U (\mu_A(y) \wedge \mu_B(y))/y$$

Similar notation of intersection is "and"; hence,

$$A \text{ and } B \triangleq A \cap B$$

Union Operation

The union of two sets A and B is denoted as $A \cup B$ and is defined by:

$$A \cup B \triangleq \int_U (\mu_A(y) \vee \mu_B(y))/y$$

The similar notation of union is "or"; thus,

$$A \text{ or } B \triangleq A \cup B$$

Complement Operation

The complement of A denoted by $\neg A$ is defined as:

$$\neg A \triangleq \int_U (1 - \mu_A(y))/y$$

The complement operation corresponds to negation. Thus if y is a member in a particular set, then not y is illustrated as $\neg y$.

Product Operation

The product operation of two sets A and B denoted by AB is defined by:

$$AB \triangleq \int_U \mu_A(y) \times \mu_B(y)/y$$

Concentration, Dilation, Contrast Intensification Operations

Concentration, Dilation and Contrast Intensification operations are the special cases of a generalized equation of a set A with the power set α, denoted A^α. The definition of A^α when α is any positive number is identified as:

$$A^\alpha \triangleq \int_U (\mu_A(y))^\alpha/y$$

Similarity when α is a nonnegative real number, A^α is defined by:

$$\alpha A \triangleq \int_U \alpha\mu_A(y)/y$$

Due to the definition of A^α, the definition for Concentration operation is represented by:

$$Con(A) \triangleq A^2$$

The definition for Dilation operation is specified by:

$$Dil(A) \triangleq A^{0.5}$$

The Contrast Intensification operation is defined by:

$$INT(A) \triangleq \begin{cases} 2A^2 \text{ for } 0 \leq \mu_A(y) \leq 0.5 \\ \neg 2(\neg A)^2 \text{ for } 0.5 < \mu_A(y) \leq 1 \end{cases}$$

Fuzzification Operation

Fuzzification operation is the complement of intensification. It is defined by the subsequent equation:

$$FUZZ(A) \triangleq \begin{cases} \sqrt{A/2} \quad \text{for} \quad \mu_A(y) \leq 0.5 \\ 1 - \sqrt{(1-A)/2} \quad \text{for} \quad \mu_A(y) > 0.5 \end{cases}$$

More complicated operations for fuzzy sets are constituted based on fundamental operations above. The meaning of former operations is to capture the vagueness in linguistic. For example, if somebody says "He is very tall", the "very" term in here adequately translates to concentration operation. In contrast, dilation operation may represent for "slightly" or "less" terms in semantics. If in the situation when any membership values which are less than 0.5 needs to be vanished, meanwhile other membership with values greater than 0.5 need to be lifted up, the contrast

intensification operation is suitable [21]. One advantage of these operations is that they can be accumulated depending on the intensive level of linguistics.

With its progressive properties and extended operations, fuzzy systems and fuzzy sets have applied and achieved many successes in many application domains such as control problems, image processing, data processing and computer vision.

2.2 CLUSTERING

2.2.1 Basic concepts

Data clustering is a computer–assisted process of exploring a relationship or analyzing hidden information of enormous sets of collected data. Data clustering belongs to "unsupervised learning" techniques in which assessible information for clustering algorithms is just data set, no guidance is needed [5], [14].

The history of clustering discipline is relatively short, starting at around 1800s. The original method was invented and applied in ecology field and by some Poland scientists who did an experiment on grouping various species in specific spots [18]. The publication of their paper attracted interests of researchers in the emerging field. Over later few years, with the appearance of several novel clustering techniques, data clustering becomes a promising method for extrating knowledge from data. It is applied in various fields such as pattern recognition, medicine, spatial image analysis, and so on [18]. In this small section, all basic concepts of data clustering

will be recalled.

Even though, the growing up of research activities in data clustering is remarkable; basically, general purpose of data clustering is still to organize observed data set into some meaningful structures [16]. According to obtained structures, hidden information or patterns inside the data can be concluded. In other words, data clustering tries to observe/use the concept of similarity between objects to determine their belonging into the same group (or cluster). Objects in same clusters share important properties with each other [5]. Briefly, data clustering is to find structures in which objects are closely-packed [35].

Clustering methods are classified into 2 main categories in terms of their similar behaviours [16]:

- Partitioning approaches:
  - The commonality of these algorithms is to construct a number of partitions depending on how to specify an integer k provided by users. In this case, k represents for the number of clusters or partitions which must be returned after algorithms converge.
  - Both algorithms use a set of criteria in order to ensure the convergence of algorithms. One popular criterion which is usually picked is a squared error function. Basically, square error function minimizes the error between cluster centroids and points assigned into that cluster while maximize the

distance among clusters defined as follow [16]:

$$e^2(X, L) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \|x_i^{(j)} - c_j\|^2$$

where $x_i^{(j)}$ is the i^{th} sample belonging to the j^{th} cluster,

k is number of clusters,

c_j is the centroid of j^{th} cluster in the cluster sets L,

n_j is the number of points in a cluster.

- Hierarchical approaches:

  This method can be understood as building a tree (or a dendrogram) for input data in which each object(point) of the data set belongs to a node of the tree. Each similarity level in the dendrogram stands for partitions of clusters [16]. The idea of having a dendrogram for input data is very fruitful, especially for applications that need generalizing inferences about input data.

  There are two styles of hierarchical methods:

  - Bottom - up (Agglomerative): This algorithm starts with one object in the input data and then merge them with a different level of similarity until the whole data set is achieved.
  - Top - down (Divisive): This algorithm is vice - versa with the above method. It starts with the data set and then splits them until each data point belongs to a node of the tree.

Besides two main methods mentioned above, many new clustering techniques have been studied and developed such as grid–based method, density–based method [5]. Each method mentioned above is practical in a specific application field, however, both of them form very useful tools to help people in analyzing data, and coming up with new insights about available data.

Referring to above approaches, distance measure acts as a key to calculate the similarity of objects in a cluster. Therefore, distance measure is a very important component that affects significantly the accuracy of clustering. Minkowski metric is a preferred distance metric to choose because of its flexibility. Minkowski metric is computed as:

$$d(x_i, y_i) = \sqrt[p]{\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)}$$

Looking at the formula, Euclidean distance can be considered as a special case with $p = 2$. Besides, Hamming distance is another popular distance which is quite commonly used in clustering algorithms. An appropriate distance metric is selected heuristically or depend on the user assumptions about the data.

Below here are benefits of clustering methods:

- According to data representation, structures of the observed data can be exploited without any explanations of the data.
- Clustering approaches both have simple mathematical presentations; there-

fore, it is relatively simple for researchers to implement them.

- Because of simple algorithms, the computational cost is low and the performance of algorithms is robust [35].
- Human-beings are good cluster seekers for low dimensional data set, for example 2 or at most 3 dimensions, but they can not deal with more than 3 dimensions. Data clustering techniques are consistent when facing with multi-dimensional data sets; therefore, the reliability of outcomes is high.
- There are various options in selecting the distance metric and algorithms. Hence, each algorithm can be applied for a specific application with substantially successful results.
- Clustering does not have identifiers associated with each object in the data set, hence, the process of collecting data is fast and straighforward approach.

In spite of the fact that clustering methods bring beneficial, they still have some drawbacks:

- Majority of clustering algorithms start with random initialization of points as cluster centroids, hence, experiments will produce different clusters in each time the clustering algorithm is run. It forces data analysts to spend time on studying the multiple results in order to conclude which is the best cluster results for the data set.
- The interpretation of the results is context–dependent.

- Another issue is cluster validation. Clustering methods both depend on the fix integer k in terms of number of clusters. Hence, how many clusters are good for the data set is another topic of clustering approaches [35].

- Clustering techniques are sensitive to noise and outliners [35].

Regardless of all disadvantages, clustering methods still attract researchers' interests and aim many accomplishments in predicting novel trends, patterns or behaviours in various fields as pattern recognition, text mining in for website, spatial data analysis and so on.

Two clustering algorithms, hard – Cmeans and fuzzy – Cmeans will be introduced explicitly in the next section.

2.2.2 Clustering Algorithms

2.2.2.1 Hard–Cmeans Clustering

Hard cmeans clustering is a fundamental clustering technique in partitioning method family (MacQueen, 1967). It is preferable for human–data analyzers because of its simple implementation, robust performance, and memory–efficiency.

Hard cmeans algorithm has following properties [14]:

- Each object in the data set is assigned precisely to one group; mathematically,

it can be defined by using crisp number representation $X = \{0, 1\}$ as:

$$\chi_{A_i}(x_j) = \begin{cases} 1 & \text{if } x_j \in A_i; \\ 0 & \text{if } x_j \notin A_i. \end{cases} \tag{2.1}$$

where χ is identified as a function representing the belonging of a point x_j to one of the cluster set $A_i, \{i = 1, \cdots, c\}$.

- A number of clusters must be prior–selected. Moreover, the number of clusters must be greater than 2 but smaller than the number of points in the data set. In particular, c is identified as the number of clusters, P is defined as number of points of the data set, hence, $2 \leq c < P$.
- Each cluster must be a non–empty cluster and it cannot contain all the data points.

The aim of this algorithm is to group a data set Z into homogenous and distinct groups $A_i, \{i = 1 \cdots, c\}$. Objects in the same cluster must be close to each other and far from other clusters. More specifically, the objective is to find the best centroid and the allocation of data points such that the distance between them is minimized. Mathematically, the above sentence is formulated as:

$$J(Z, V) = \sum_{i=1}^{c} \sum_{j \in A_i} d(j, v_i) \tag{2.2}$$

where V is defined as the vector of centroid,

Z is the data set to be clustered,

$d(j, v_i)$ is a suitable distance metric between the data sample x_j and i^{th} cluster centroid v_i:

$$d(j, v_i) = d(x_j - v_i) = \|x_j - v_i\|^2 \tag{2.3}$$

The centroid v_i in the first iteration of the algorithm is an arbitrary selected point from the data set. After the first partitioning, the centroid v_i in the k^{th} iteration is recalculated by averaging all points in the same cluster as:

$$v_i^{(k)} = \frac{1}{G} \sum_{x_j \in A_i} x_j \tag{2.4}$$

where G represents for total number of samples in the cluster A_i.

The algorithm stops if the best optimum combination of (U^*, V^*) is found which minimize the equation (2.2). Differently, the algorithm converges if the difference of $J(Z, V)$ in two consecutive iteration smaller than a pre–selected ε.

The algorithm is summarized :

- Step 1: Initialize the centroids randomly by selecting number of data points in the data set corresponding with the given number of clusters.
- Step 2:
  - Assign each data point to the closest cluster centre by using the formula (2.3).

- Calculate $J(Z, V)$ in the iteration $k+1$. If $\|J^{k+1} - J^k\| < \varepsilon$ then stop, else go to step 3.

- Step 3: Compute new centroids using the equation (2.4). Go back to step 2.

Hard – cmeans algorithm is suitable for data sets that include clusters that have similar shapes and same size. Otherwise, hard–cmeans might converge at local minima [35]; as the result, the global optimization of distance does not reflect the right partitions and the shape of the cluster is not compact.

In order to overcome this drawback, the concept of fuzzy sets is applied into hard–cmeans and enhanced it to be fuzzy–Cmeans which is introduced in the next section.

2.2.2.2 Fuzzy–Cmeans Clustering

Fuzzy–Cmeans (FCM) clustering [4] is first devised by Dunn in 1973 and then improved by Bezdek in 1981. This algorithm is an extension of classical hard–cmeans by introducing fuzzy memberships associated with each data point. Like hard cmeans clustering, a number of cluster c must be prior – known. The number of clusters must be greater or equal to 2 but smaller than number of data points.

Besides, FCM algorithms possess other properties which are considered as big improvements comparing with hard cmeans:

- FCM technique introduces the fuzzy concept into the algorithm. Each data point is associated with a vector of fuzzy memberships that show the degree of belonging of the point to each cluster. Each membership value in here is a number from a set $X = [0, 1]$. In other words, objects in fuzzy cmeans algorithms can belong to more than one cluster with different degrees of membership.

- Instead of having hard boundaries like in the hard–cmeans clustering method, the FCM supports the concept of soft boundaries. In the FCM technique, the shape and size of cluster can be flexible based on the degrees of membership. The larger of degrees of memberships, the closer of the point to its centroid, clusters are more compact. Hence, inferences about the quality of clusters just depend on selected membership value.

Again, the aim of the FCM algorithm is to find the group of fuzzy partitions of a data set X. More formally, the objective of the algorithms is to minimize the total distance between data points and its cluster centroid. In addition, due to the second property alluded above; another task is to maximize the degree of memberships [26]. Imposed on two indications above, a generic formula is [19]:

$$J(U, V; X) = \sum_{i=1}^{N} \sum_{j=1}^{c} u_{ij}^{m} \|x_i - v_j\|^2 \tag{2.5}$$

where U is the fuzzy partition matrix corresponding with the data set X which must

satisfy a condition:

$$\sum_{j=1}^{c} u_{ij} = 1 \tag{2.6}$$

$V = (v_1, v_2, \cdots, v_c)$ is the vector of cluster centroids.

$m, (1 \leq m < \infty)$ is the fuzziness parameter which controls the degree of membership values. Differently, it controls the fuzzy partition of the results. The larger m is, the fuzzier the memberships. Normally, $m = 2$ is most common choice.

In order to achieve the optimal solution, the algorithm has to pass through some iterative steps. It leads to the updates for both membership matrix and centroids. The formulas that govern the updates are:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - v_j\|}{\|x_i - v_k\|} \right)^{\frac{2}{m-1}}} \tag{2.7}$$

and

$$v_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{2.8}$$

The iteration stops when the difference of fuzzy matrix in two consecutive iterations is smaller than a pre - selected ε or a pre–defined number of iterations is reached.

Ultimately, the algorithm composes all following steps:

- Step 1: Initialize randomly fuzzy membership values for each data point corresponding with the number of cluster c. The fuzzy membership values must

be under the constrain (2.6).

- Step 2: Calculate the vector centroids associated with the fuzzy membership matrix which are just generated above using the formula (2.8)
- Step 3: Compute the disimilariry between centroids and data points using the equation (2.5). If $\|J^{k+1} - J^k\| \leq \varepsilon$ then stop; else go to step 4.
- Step 4: Update the membership value matrix and centroids by following the formulas (2.7) and (2.8).

2.2.3 Clustering – based Classification

In this section, fundamental notions about classification and clustering–based classification methods are reviewed.

Classification belongs to the category of "supervised learning" methods in which each object in training data set accompanies with a label that shows its belonging to a specified class [14]. More formally, the label acts as a teacher (or guidance) to drive the output.

As a generic perspective, there are two goals while using classification techniques. The first target is to build a classifier to classify the data based on its past examples. The obtained classifier can be used to classify new, unknown data. The second objective while classifying data is to construct a model which serves as a predictor to anticipate future events, detect abnormalities in the data, or find missing

variables [35].

In the classification phase, construction of classifiers are based on two sets: a training sets and 10 fold–cross validation. Normally, the given data set is divided into 10 equivalent segments. A popular recommendation for all classification algorithms is to take $\frac{9}{10}$ segments to be a training set. The training set serves as past samples in order to build a classifier (or a model). An achieved model (or classifier) will be applied on 10 fold-cross validation sets (in which each instance is taken out its label) in order to validate how good the model (or a classifier) is. If the classifier can produce exact labels as the desired labels which are assigned previously for 10–fold cross validation sets, the classifier have high accuracy and can provide precise predictions or good classification for new, unknown patterns. There are many algorithms that can be used for classification such as k-NN (k–nearest neighbor), decision tree association rules which gain a lot of popularity in numerous fields as text mining, pattern recognition and so on.

While clustering and classification methodologies take a pretty good care of labeled and unlabeled data separately, respectively; both indicated methods might not perform well when coping with the training set included both labeled–unlabeled entities. Clustering–based classification method (CBC) is proposed as a solution to handle labeled–unlabeled instances in training set [36]. CBC technique involves two main steps: clustering and classification; however, the way to treat the obtained knowledge from clustering methods dues to application domains and user demands.

For example, according to a paper of Hua-Jun Zeng [36], the authors proposed CBC which is applied in text mining. The problem of text mining is there are the abundant available sources of unlabeled text documents; meanwhile, there are only small sources of labeled text documents. In order to satisfy its properties and increase the accuracy of the application, CBC is proposed as follow:

- Clustering techniques is applied to group the training set which involves both unlabeled and labeled instances into clusters. Based on the similarity of features among instances in the same cluster, unlabeled instances are assigned the same label as labeled data points.
- Classification step: use the expanded labeled training set in order to construct a classifier.

The accuracy of classifying text documents after employing CBC technique is improved significantly.

Another example illustrating the variance of CBC technique in different contexts is mentioned in the paper of Ray H.Hashemi and Mahmood Bahar [12]. In this paper, CBC technique is described as:

- Clustering step: a regular clustering approach is applied on the training set which contains unlabeled data points, resulting in seperated clusters. Then, each obtained cluster is assigned a specific label; differently all objects belonged to each cluster correspond with a specific label.

- Classification step: a new data set is already generated, each object accompanied with a label. Any classification method is picked to train the data set in order to get a classifier or model based on the new generated labeled training set.

Two examples above shows that CBC technique varies based on how information from clustering step will be treated and utilized in each application. Our approach is also a derivation from CBC approach but information from clustering step is extracted and exploited differently which will be illustrated in a great detail in chapter 4.

2.3 EVIDENCE THEORY

2.3.1 Basic Concepts

Originally, uncertainty is admitted under a form of the idea of chance or the degree of belief [23]. The idea of chance (or the degree of belief) is a subjective measure to evaluate the certain occurrence of an event according to our empirical observations or experience. Under the mathematic consideration, the idea of chance and the degree of belief is combined under a name called probability. Probability measure uses the number in the range $[0, 1]$ to assign for the certain occurence of one event. The higher likelihood measure in the range $[0, 1]$, the greater chance that event will happen. The likelihood value 0 stands for totally fault or not happening at all, on the other hand, 1 stands for totally true or sure happening [20].

Traditionally, the foundation for probability methods is represented by the Bayesian theorem [20]. The aim of Bayesian theorem is to calculate a conditional probability based on two events from another conditional probability. Mathematically, Bayesian theorem is identified as follow:

$$Pr(A|B) \propto Pr(B|A) \times Pr(A)$$

where $Pr(A)$ is the prior – probability of A

$Pr(B|A)$ is the conditional probability of B, given A.

$Pr(A|B)$ is the conditional probability of A, given B. It is also called as posterior probability because it depends on the value of B

Despite of the fact that probability, especially Bayesian theorem is the appropriate method to deal with numeric probability, it still has its own problems:

- Probabilistic methods requires a subjective measure to assess uncertainty regarding to a specific phenomena; however, it is already a big issue for people because each person has different points of view while observing even the same event. Obviously, due to different viewpoints, the probability measure will come up with different results.
- Probability requires having available information about all events but in some cases, the information is not available. In this case, all events are assigned equally likelihood.

- There are two concepts: belief or disbelief associated with probability measures. In other words, regarding to the same event, belief represents for the certain occurrence (or true), meanwhile, disbelief represents for nonoccurrence at all (or fault) [20]. However, this clause is not right somehow because disbelief doens't mean the event has to be totally nonoccured. It might refer to a lacks of knowledge to conclude whether that event totally doesn't happed or it might commit to an unknown part.

- Probability is not versatile for representing or combining multiple sources of different types of belief.

Evidence theory is introduced as a remedy to overcome most of probability limitations mentioned above. Evidence theory was proposed by Shafer in 1976, and was an extension of the work done by Dempster in 1967 [23]. Dempster– Shafer theory is considered as a generalization of Bayesian theory of subjective probability because not only it maintains probability properties but also it refines and offers extra features that traditional probability doesn't have:

- In constrast to traditional probability approaches which handle only one possible set of events, Dempster–Shafer theory allows for working with multiple sets of events (or a set of propositions). In other words, the theory deals with the frame of discernment $\Theta = \{\theta_1, \theta_2, \cdots\}$ in which $\{\theta_1, \theta_2, \cdots\}$ are possible worlds or possible outcomes of what an agent considers possible. Thus, the prepositions of interests are in to one-to-one corresponding with the subsets

of Θ and its power 2^{Θ} represents the set of all prepositions of interest [23].

- The model does not need any assumption or any prior information about the sets; in particular, the distribution, prior-probability, etc. This feature gives the model a capability to deal with multiple levels of precision regarding information.

- The framework still remains the complement rule for belief and disbelief measure. However, it extends the former rule by introducing the ignorance concept [20]. The ignorance concept is understood that the agent should concern about what it wants to know and ignore details that it does not have any information about it (or doesn't know anything about it at all). According to this idea, the disbelief part now might commit to an unknown part or/and non–happenning part.

- The heart of Dempster–Shafer theory provides a mechanism for uncertainty a capability to deal with multiple sources of different types of belief by using the combination rules. The degree of belief for a particular event is enhanced aided by applications of Dempster–Shafer combination rules.

According to its advanced characteristics, Dempster–Shafer theory brings innovative and fertile perspectives for researchers while dealing with uncertainty.

There are three important functions in Dempster–Shafer theory: **the basic probability assignment function (b.p.a), the Belief function (Bel) and the**

Plausibility function (Pl).

The basic probability assignment function

The basic probability assignment function (b.p.a) is the basic function to evaluate the degree of uncertainty of a proposition. It has another name as a probability mass function (define as m). The b.p.a is a mapping from a power set of all possible propositions to the interval between 0 and 1 which contains following characteristics:

- The value of the b.p.a for the given set A (represents as $m(A)$) stands for the available evidence which supports set A and only set A and it doesn't give out any indications about the support evidence for any subsets in A. Any further evidence about the subsets of A would be represented by another mass function.
- The value 0 is assigned to an empty set.
- The summation of mass functions of all subsets of the power sets is 1.

Formally, the description of mass function is represented as:

$$m : 2^{\Theta} \rightarrow [0, 1]$$

$$m(\emptyset) = 0$$

$$\sum_{A \subset \Theta} m(A) = 1$$

Notion of Belief

Even though a mass is a good representation of the belief associated with a single subset; some proofs indicated that a mass is not sufficient to represent a probability in the sense of classical probability [20]. The mass does not contain adequate belief to represent the total belief supporting for one subset of Θ. Fortunately, Dempster–Shafer theory can support this issue. Evidence theory defines the total belief committed to a set A by summing up all the probability masses from all proper subsets which is contained in A:

$$Bel(A) = \sum_{B \subset A} m(B)$$

Based on the belief definition, the concept of ignorance is also addressed. So far, we know that the belief function represents the belief assigned to a particular subset A in Θ, it means that all other subsets are ignored and there is no knowledge about them. It is the definition of ignorance concept. This concept allows us to assign the belief value equal to $1 - Bel(A)$ to any subsets in Θ where we lack of knowledge or do not know anything about that. This notion is clarified by an example as:

Suppose that we have a proposition represented by a subset A of Θ. Its belief values is identified as b, where b is in the interval $[0, 1]$. More specifically, we have $Bel(A) = b$. Further, we know nothing rather than the information for A. The

ignorance concept for Θ is represented as $Bel(\Theta) = 1 - b$.

Belief Function

According to Dempster– Shafer theory, rather than combining all the possible mass together in order to have the total belief, Dempster– Shafer provides another approach to find a belief function without using masses. The approach is described as follow:

A function $Bel : 2^{\Theta} \rightarrow [0, 1]$ is called as belief function if and only if satisfies the following conditions:

$$Bel(\emptyset) = 0 \tag{2.9}$$

$$Bel(\Theta) = 1 \tag{2.10}$$

For every positive integer n and every collection $\{A_1, \cdots, A_n\}$ of subsets Θ,

$$Bel(A_1 \cup \cdots \cup A_n) \geq \sum_{\substack{I \subset \{1,\cdots,n\} \\ I \neq \emptyset}} (-1)^{|I|+1} Bel(\bigcap_{i \in I} A_i) \tag{2.11}$$

Furthermore, it is also possible to obtain the basic probability assignment from the Belief function by using the inverse function:

$$m(A) = \sum_{B \subset A} (-1)^{|A-B|} Bel(B)$$

where $|A - B|$ is the difference of the cardinality of the two sets.

Plausibility Function

Plausibility function is defined as a measure of the extent that A is plausible[23]. $Pl(A)$ is obtained as follows:

$$Pl(A) = \sum_{B \cap A \neq 0} m(B)$$

where $B \in 2^{\Theta}$.

It also might be written as:

$$Pl(A) = 1 - Bel(\neg A)$$

where $Bel(\neg A)$ represents for the extent that one believes in its negation $\neg A$.

The interval $[Bel(A), Pl(A)]$ is regarded as the lower and upper probability to A. In addition, the interval illustrates the ignorance measure regarding to the set A. The value of ignorance measure might vary from 0 to 1. When the $Bel(A) = 0$, then $Pl(A) = 1$, it means that there is no mass committed to A and also any of its subsets, but also no mass is assigned to it $\neg A$.

2.3.2 Dempster – Shafer Combination Rules

To reiterate, the most significant contribution of Dempster – Shafer evidence theory is that it allows for combining different kinds of evidence from different

sources. The obtained results by integrating multiple evidences together could increase beliefs for a specific problem.

As Karl Sentz and Scott Ferson indicated in [24], there are four different types of evidence:

- **Consonant Evidence:** It can be represented as a set of nested subsets in which a smaller subset is totally inside a bigger subset. It can be illustrated in Figure 2.1

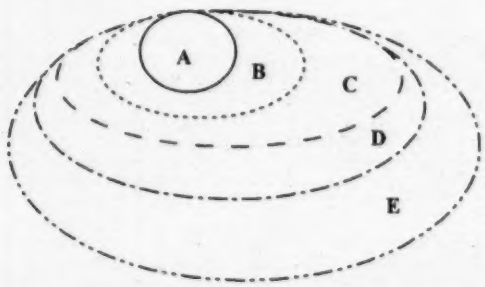

Figure 2.1: Consonant Evidence obtained from multiple sources

- **Consistent Evidence:** There is only at least one common subset for all multiple subsets collected from multiple sources. The figure 2.2 below represents the idea of consistent evidence obtained from multiple sources.

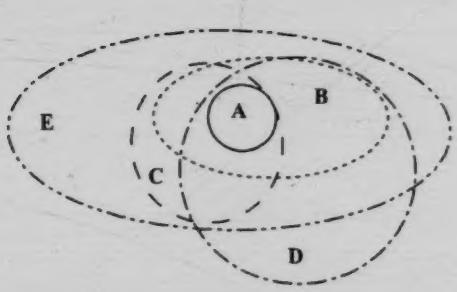

Figure 2.2: Consistent Evidence obtained from multiple sources

- **Abitrary Evidence:** In contrast with consistent evidecence, arbitrary evidence implies that there is no common subset among a set of multiple subsets, although, there are many common subsets between two subsets in a set of multiple subsets. It is displayed as in Figure 2.3:

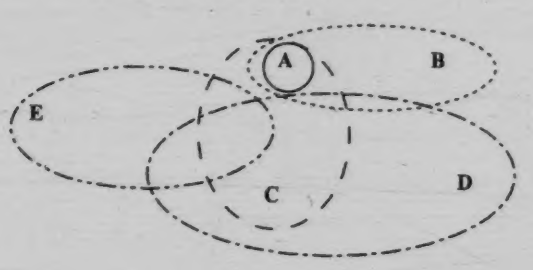

Figure 2.3: Arbitrary Evidence obtained from multiple sources

- **Disjoint Evidence:** All subsets are seperated from each other which are illustrated in Figure 2.4:

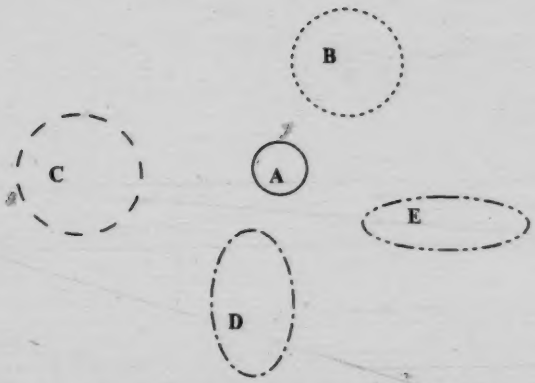

Figure 2.4: Disjoint Evidence obtained from multiple sources

Based on the Dempter – Shafer definition, the general rule for combining two belief functions, regarless the types of evidence, usually is represented through the b.p.a (or masses). Two belief functions must be defined in the same frame Θ and the aggregation between their masses over focal elements must satisfy:

$$\sum_{\substack{i,j \\ A_i \cap B_j = \emptyset}} m_1(A_i) m_2(B_j) < 1$$

where A_i, B_j are focal elements corresponding with each belief function, and m_1, m_2 are the b.p.a for each focal element.

The core belief function aggregated from two masses above is yielded as:

$$m(A) = \frac{\sum_{\substack{i,j \\ A_i \cap B_j = A}} m_1(A_i) m_2(B_j)}{1 - \sum_{\substack{i,j \\ A_i \cap B_j = \emptyset}} m_1(A_i) m_2(B_j)}$$

Obviously, the general combination rule above is valid for any types of evi-

dence. Furthurmore, it also applies for combining any number of belief functions by accumulating pairwise belief functions together.

However, the combination rules would be more appliable if they have specific combination formulas for each types of evidence. Dempster – Shafer theory addresses this problem by defining three combination rules regarding to four distinct types of evidence explained above:

- **Combination rule for homogeneous evidence:** This rule supports for the combination of evidences which point to one specific subset. The formula for this combination is specified:

$$m(A) = 1 - (1 - s_1)(1 - s_2) \text{ and } m(\Theta) = (1 - s_1)(1 - s_2)$$

where s_1, s_2 are probability numbers supporting subset A from two different sources.

- **Combination rule for heterogeneous evidence:** The body of evidences in here points to different subsets. Moreover, the intersection between subsets does not equal 0. The equation is defined as:

$$m(A) = s_1(1 - s_2), \quad m(B) = s_2(1 - s_1),$$

$$m(\Theta) = (1 - s_1)(1 - s_2), \text{ and } m(A \cap B) = s_1 s_2$$

where s_1, s_2 support for each subset A, B respectively.

- **Combination rule for conflict evidence:** The body of evidences in this case is not only heterogenous but also conflicting with each other. It implies that the intersection between two subsets equals to 0. The equation is represented as:

$$m(A) = \frac{s_1(1-s_2)}{1-s_1s_2}, \quad m(B) = \frac{s_2(1-s_1)}{1-s_1s_2}$$

$$m(\Theta) = \frac{(1-s_1)(1-s_2)}{1-s_1s_2}$$

where s_1, s_2 are defined as previously.

These formulas can be generalized while coping with any number of subsets and basic probability assignment functions.

CHAPTER 3

RELATED WORKS

This chapter focuses on reviewing some applications of Dempster – Shafer theory to combine results obtained from multiple base–level classifiers which has been reported in a number of research publications over the period of last several years. The history of research, key ideas as well as summaries of general experimental results in each approach are briefly mentioned and summarized.

One of the first works in this area is dated back to 1988 when a method was proposed to transform distance measures of the different base-level classifiers into evidence [17]. Once the distances between learning data points and a number of reference points have been calculated, they were used for evaluation of basic probability assignment values, and later the Dempsters combination rule (eq. 18) was used to combine these basic probability assignment values. Different proximity mea-

sures between a reference vector and a classifiers output vector were investigated by Rogova in [22]. The measures with the highest classification accuracy were further transformed into evidences, which were merged in order to obtain an overall confidence measure for each category. The merge of evidences was performed according to Dempster-Shafer rule of combination.

The work reported by Denoeux in [8] focuses on the application of Dempster–Shafer evidence theory to evaluate the classification results by combining results obtained from the k–nearest neighbor classification method. The construction of a new classification procedure is summarized explicitly as follows: By applying the k–NN classification methods on the training set, a set of nearest neighbors for a sample under consideration were achieved. Each class associated with each neighbor was regarded as a piece of evidence supporting the class that a sample belongs to. Once pieces of evidence were obtained from the k-NN classification method, the values of basic probability assignments were determined as the distances between a sample under consideration and its neighbors. The bpas then were merged by means of Dempster–Shafer combination rules to yield final classification results for a questionable sample. The significant improvement of the proposed method comparing with other methods is that it provided the global treatments for issues such as ambiguity and distance rejection, imperfect knowledge regarding the class memberships of training patterns based on the concept of Belief and the Plausibility Functions. Simulations on both synthetic data sets and real data set were revealed

the better effectiveness of the proposed approach as compared to the voting and distance–weighted k–NN rules in various aspects. For example, the estimated error rate related to the decision boundary and distance rejection issue was 0.084 for the proposed method, against 0.089 for the voting 9-kNN.

Another work also reported by Denoeux where the author proposed another method of calculating degree of support [9]. In this work, the procedure of assigning a class to a sample was determined by extending the concept of finding nearest neighbors in the training set to a limited number of representative patterns or *prototypes*. Each prototype is assumed to possess *a degree of membership* to each class with the constrain $\sum_{q=1}^{M} u_q^i = 1$ where M represents the total number of classes in the training set. Full membership of a prototype to one class was considered as the special case where $u_q^i = 1$ for some q and $u_l^i = 0$ for $l \neq q$. The basic probability assignments were calculated using distances between a point and a limited number of prototypes, and the degrees of membership of these prototypes to each class. A very interesting element of this work is related to variations in a selection of a "winning" class. The approach presented there considered possible consequences of different actions, in particular a rejection of ambiguous patterns. Once they were quantified they were included in calculations of risk relative to the pignistic probability distribution [10]. A very interesting approach to evaluate evidences (bbas) was presented by Al-Ani and Deriche in [2]. The method was based on a concept of tuning basic probability assignments during a training process so that the overall

mean square error of an ensemble classifier is minimized. Several experiments were simulated on real data sets to demonstrate various aspects of the new classifier. The first experiment was simulated on the famous IRIS data set to illustrate the form of the output of the model, as well as the decision regions. The second experiment was simulated on two data sets which are the pheneme recognition data and the forensic glass data. Results obtained from the new classifier were compared with some past classifiers to assess its performance. Overall, the proposed classifier gave out better classification comparing to other past models and it allows for efficient rejection of outliners. The third experiment focused on simulating a data fusion application to prove the robustness of the new classifier. Results combined from two classifiers on the framework of Dempster–Shafer evidence theory were compared with results combined on the Bayesian probability framework.

A method of combining classifiers with different sets of classes was investigated by Ahmadzadeh and Petrou in [1]. In this paper, the proposed approach focused on combining results achieved from two classifiers which were Bayesian network classifier and fuzzy logic–based classifier aided by evidence–theoretical framework. Each classifier mentioned in this paper is regarded as a base–level classifier. The distinction of the proposed method with other past research was that each base-level classifier employed in the paper provided a prediction into a different set of classes. As a brief recall, one of the condition to assure Dempster–Shafer theory to work properly is all sources should have the same frame of discernment. However, since

each classifier produced a prediction on different set of classes, it means that results obtained from each classifier did not have the same frame of discerment. As the obvious consequence, Demspter–Shafer rules of combination could not perform in this case. This problem was solved using superset of finer classes which was defined as the union of the number of classes obtained from base–level classifiers. Superset of finer classes then could be combined to produce classes according to any base-level classifier. The application of the Dempster-Shafer theory provided a way of doing that via taking into account relative reliability of base-level classifiers. Each base-level classifier induced a single belief structure, and values of basic probability assignments were calculated based on output of a base-level classifier and its reliability measure. The effectiveness of the proposed approach is tested to predict soil erosion problem as compared to fuzzy classifier alone. The corresponding results showed that six out of 9 testing sites were correct classified regarding to the new method, against five out of nine sites for fuzzy classifier.

CHAPTER 4

APPLICATION OF EVIDENCE THEORY TO CLUSTERING-BASED CLASSIFICATION

4.1 OVERVIEW

In the previous sections basic elements of a few data processing and analyzing methods such as clustering and fuzzy clustering, clustering-based classification, and kNN are described. These methods are commonly used for building prediction/estimation models.

It should be indicated that each of these methods treats data differently, each of them "see" and process data in a different way, and at a different level. For example, kNN takes into consideration local aspects of data points that are close to the point being classified; clustering–based classification, on the other hand, "looks at" all data points at the same time, and classifies a data point based on

comprising of this data point with groups of data points. In many cases these methods provide relatively good predictions. However, their performance depends on nature/character of data used for building classifiers.

The aspect that is not covered in any of these methods, and other most commonly used ones, is the issue of a measure of belief in the obtained results/classifications. Classifiers seem to be "very sure" about their predictions. Some of the exceptions of these statement can be seen in the case of ensemble of classifiers where user can "see inside" and obtain more information about how the final results were obtained - how many basic classifiers (basic classifier is a single classifier in the ensemble) identified given class, and based on that "figure out" his/her confidence in the result.

In this chapter, a new approach that tries to identify both mentioned above issues - taking into consideration different ways of "looking at" the data, and obtaining some indicator about confidence in obtained results is introduced. This approach is a combination of two classification methods with elements of evidence theory.

Let us start the description of the method by identifying two classification methods that constitute the components of the proposed approach. One of them is kNN, and the other is clustering–based classification.

The kNN method has been selected because of its emphasis on the **local** aspects of data distribution. This method provides prediction based on comparison

of a new data point with its closest neighbors. It does not see a "big picture". No matter what kind of points are outside "circle of neighbors" - kNN method provides the results using only on a few data points. So, the advantage of the method is that is tries to look at "peers" (close neighbors), and assumes that majority of points will be able to "decide" about the result. The disadvantage is that the method "looses" the view at the level of collections or groups of points, and does not gain anything from global distribution of points.

The second method used in the approach is the method based on cluster-based classification. This method provides aspects related to a **global** view. Our utilization of the approach focuses on a global "trends" in data distribution. The advantages and disadvantages of this method are reciprocal to the disadvantages and advantages of kNN method. In this case, the classification process of a new data point takes into consideration a location of a new data point in reference to centroids of data clusters.

In the proposed approach, the element that "binds" these two methods together is evidence theory. It is used to combine classification results obtained from each classifier. The processes of deriving a result used in each method are also altered by application of elements of evidence theory. The details describing the application of evidence theory in each method are presented in Sections 4.2 and 4.3 respectively.

4.2 LOCAL-BASED CLASSIFICATION

4.2.1 Concept

The local-based classification is performed using k Nearest Neighbors (k-NN) method [6] [7]. The principle of this methods is to select a number of points – "k" – that are closely located to a new data point that is being classified. A category of this new point is identified based on categories of neighbor points. This process is performed using the voting" approach or majority" approach – a category of the new point is related to classes of neighbors. The indicated approach is not very accurate if there does not exists any dominant class among neighbors. Additionally, the issue of closeness of the neighbors to the point being classified is not taken into consideration.

In order to address the above mentioned issues, Denoeux [8] [9] [10] proposed the concept of application of elements of evidence theory to kNN method. The proposed approach follows this concept. The process of determining a class of the point being classified picks up the "k" neighbors, but the process of identifying a dominating class (among neighbors) is altered. This alternation is related to a step of calculating a belief representing an importance of contribution of a single neighbor to the overall result. In a nutshell, further the point from the point being classified lower the belief in a contribution of this point to the final result.

4.2.2 Calculation of Belief Masses

Each belief mass is an evidence supporting the statement that the point belongs to the same category to which a given neighbor belongs. The principle used for calculating beliefs is very simple – values of belief masses are calculated based on the distances between a point being classified, we will call it *thenewpoint* hereafter, and its neighbors.

The first step is to find neighbors. In this approach, the neighbors are all points located in the circle with a radius "r" and the new point as its centre. The value of "r" defines a number of neighbors. Of course, the relationship is simple – bigger "r" leads to a larger number of neighbors. Identification of neighbor means calculating distances between the new point and the points with already known categories. Euclidean distance is selected for this purpose.

The second step is to use already calculated distances to calculate individual belief masses associated with each neighbor point "p", which belongs to a category c_j, based on the following equation:

$$m_p(category = c_j) = 1 - \frac{d_p}{\sum_{i=1}^{k} d_i} \tag{4.1}$$

where $m_p(category = c_j$ represents a belief that the new point belongs to the category c_j, d_p is a distance from the new point to its neighbor p, and $\sum_{i=1}^{k} d_i$ represents the sum of distances between all neighbors and the new point. A closer look at

the equation indicates that even the furthest point, let us assume that it belongs to the category c_n, provides a non-zero evidence (belief mass) contributing to the statement that the new point belongs to the category c_n.

4.2.3 Local Beliefs

Assume that there are N different categories. The belief masses (contribution form each neighbor) calculated in the previous section are used to calculate beliefs supporting belonging of the new point to each category. all beliefs that are related to the same category are grouped. For example, suppose that points q, r and s belong to the category c_j, then the three masses $m_q(category = c_j)$, $m_r(category = c_j)$, and $m_s(category = c_j)$ are three evidences supporting belonging of the new point to the category c_j. These evidences are combined using the Dempster's combination rule for homogenous evidence (see Section .... for details). This process is repeated for every category. As the result N beliefs are obtained: $m(category = c_1)$, $m(category = c_2)$, ..., $m(category = c_j)$, ..., and $m(category = c_N)$.

These N evidences that are obtained based on combining the homogenous evidences are not the final beliefs supporting each category. The last step is to combine these evidences using the Dempster's combination rule for heterogeneous evidences (see Section ...). Once this is performed the final N values are obtained. These belief will be labeled with the label "L" (for local): $m_L(category = c_1)$, $m_L(category = c_2)$, ..., $m_L(category = c_j)$, ..., $m_L(category = c_N)$.

The beliefs $m_L(category = c_j)$ for $j = 1, 2, ..., N$ will be used in the final phase

of the proposed approach.

The described process of calculating local belief is presented in Figure 4.1.

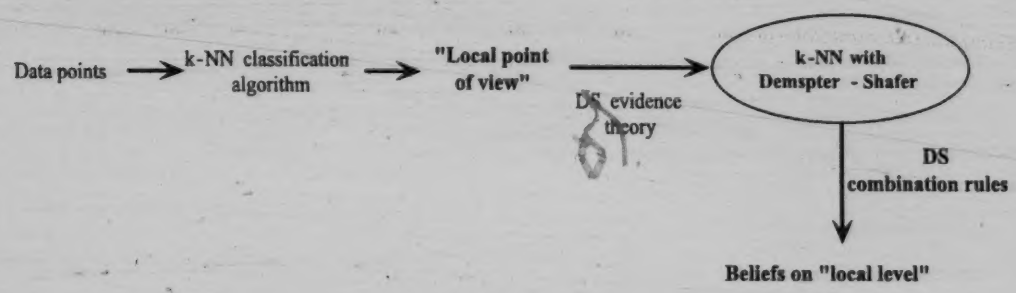

Figure 4.1: Process of obtaining local information

4.3 GLOBAL-BASED CLASSIFICATION

4.3.1 Concept

A global-based classification procedure offers a mechanism to classify new data points using a combination of a clustering algorithm and evidence theory. More specifically, information extracted from clustering-based classification (CBC) is represented in evidence-based framework as an uncertainty measurement for the assignment of a new point with respect to a specific class. The global classification technique has some similarity with CBC (Section 2.2.3). However, the difference is how the information extracted from clustering results is represented and combined.

The CBC techniques cluster a training data into distinct and compact clusters, regardless of classes the data points belong to. Herein, distances from a new data point to cluster centroids are major information obtained from the clustering

algorithms. A new data point belongs to a cluster whose distance between a point and a cluster centroid is the shortest. Then, based on the classes of data points in a cluster, a new data point is assigned to the same class as the dominant class of the cluster.

However, it is argued that an inference about the assignment of a single class to a new data point might be wrong if there is no dominant class inside a cluster. With this argument in mind, the research investigates a more rigid and reasonable technique while making an inference about a label of an unknown data point. This work still utilizes obtained clustering results, however, it generalizes them by introducing the notion of "purity".

The concept of "purity" can be defined as an alternative to evaluate the quality of each cluster based on classes associated with data points that belong to a cluster. A cluster is denoted as "pure" if it includes only points that belong to the same category. Otherwise, the cluster is described by a set of all classes that are represented by cluster data points. This type of description is easily handled within the framework of Dempster - Shafer theory. With this enhanced approach, more vivid and trustworthy information about classification is gained on global level for a new data set.

4.3.2 Calculation of Belief Masses

The first step in calculating belief masses at the "global" level is application of FCM algorithm to a training set. FCM method is preferable in this research because besides "finding" clusters it provides a membership vector with information about a degree of belonging of each data point to each cluster. FCM also supports a control of size and compactness of clusters aided by a threshold. Herein, the threshold is considered as a parameter which determines the scatter of a cluster. The higher the threshold is, the more compact and distinct cluster is and vice versa. Obviously, the value of a threshold should be selected based on applications or user demands. In research's experiments, the threshold is chosen as 0.6 to ensure clusters which are not too compact, but also not too scattered.

After applying FCM algorithm on a training set, cluster centroids and information regarding to clusters such as data points, and membership vectors are obtained. In this work, cluster centroids and information about clusters are emphasized because they are regarded as major factors that affect the process of calculating global masses in such a way that:

- Cluster centroids are employed to calculate distances among them and a new data point. It is worth noting that these distances are treated as parameters which represent relationships between a new point and clusters. The longer the distances are, the less significant relationships between a new data point and each cluster are. We propose a coefficient that represents this relationship:

$$degree(G_j) = 1 - \frac{d(G_j)}{\sum_{i=1}^{k} d(G_i)}$$

where $degree(G_j)$ represents the degree of belonging of a new data point to the cluster G_j; $d(G_j)$ is the distance between a new data point and the centroid of the cluster G_j; k is a number of clusters, and $d(G_i)$ denotes distances from a new data point to centroids of clusters G_i.

- Taking advantages of existence of classes associated with each data point in the training set, statistics about the purity of each cluster are computed. The purity $purity_{G_j}(category = c_k)$ of cluster G_j is simply defined from the percentage point of view. Ultimately, the purity of each cluster is merely the summation of all data points which carry identical class divided by the total number of points in a cluster. For instances, let's assume that there are three categories c_1, c_2, c_3 in a training set, and three clusters – G_1, G_2, G_3 – are obtained as the result of the application of FCM clustering algorithm. Further, let us assume that there are 100 points in cluster G_1, 30 of which belong to the category c_1, 30 points belong to the category c_2, and 40 to the category c_3. The statistics of purity for cluster G_1 are:

$$purity_{G_1}(category = c_1) = \frac{30}{100}$$

$$purity_{G_1}(category = c_2) = \frac{30}{100}$$

$$purity_{G_3}(category = c_3) = \frac{40}{100}$$

A coefficient $degree(G_j)$ for the cluster G_j, and a set of purities for this cluster $purity_{G_j}(category = c_k)$ (for each category) are used to calculate beliefs provided

by the cluster G_j that a new point belongs to each of the categories. This belief is computed as the product of the coefficient associated with the cluster and its purity statistic. This procedure also is named "clustering based classification with Dempster – Shafer ". For instance, if the coefficient $degree(G_j)$ for a new data point and the cluster G_j is 0.8, then, the beliefs committed to each category based on the cluster G_j are:

$$m_{G_j}(category = c_1) = 0.8 * \frac{30}{100}$$
$$m_{G_j}(category = c_2) = 0.8 * \frac{30}{100}$$
$$m_{G_j}(category = c_3) = 0.8 * \frac{40}{100}$$

4.3.3 Global Beliefs

A classification process of a new data point on global level is accomplished by global beliefs in which each belief is defined as the degree of support for a particular category. According to global beliefs, an appropriate category is assigned to a pattern under consideration on global level. In this work, global beliefs associated with a considering point are specified as a vector where the number of elements in the vector is distributed such that, the first element represents for an uncommitted part which represents for the ignorance knowledge, the rest elements shows the beliefs supporting for each category with respect to that considering point.

The calculation of global beliefs is performed by applying the Dempster–Shafer combination rules on belief masses obtained during the previous step. However, due

to the fact that there are a number of clusters, the calculation process is considered slightly complicated. Therefore, to simplify the illustration of this process, two small steps are taken into consideration:

- The first step – regarded as an intermediate step – computes a single belief that a new data point belongs to the category i based on all belief masses $m_{G_j}(category = c_i)$ supporting the category i calculated for all clusters G_j where $j = 1, 2, ..., k$. The Dempster– Shafer combination rule for homogenous evidences is used here.

- The second step – the final step – is to compute global beliefs using the Dempster–Shafer combination rule for conflict evidences. The combination rule for conflict masses is used because belief masses computing in the first step provide supports for different categories. Therefore, the rule for combining conflict masses allows to elevate global beliefs including the part committed to $\emptyset$.

The idea of computing global beliefs is illustrated explicitly in Figure 4.2. In this figure, we work on the assumption that the data points belong to three categories, and there are three clusters. Each cluster includes points all three categories inside itself.

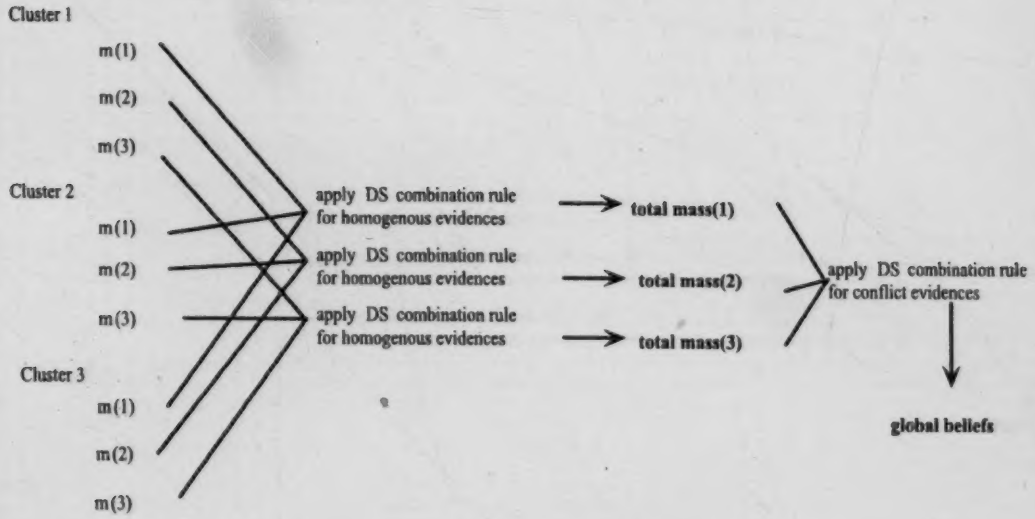

Figure 4.2: Process of calculating global beliefs

The whole process of calculating global beliefs is summarized in Figure 4.3.

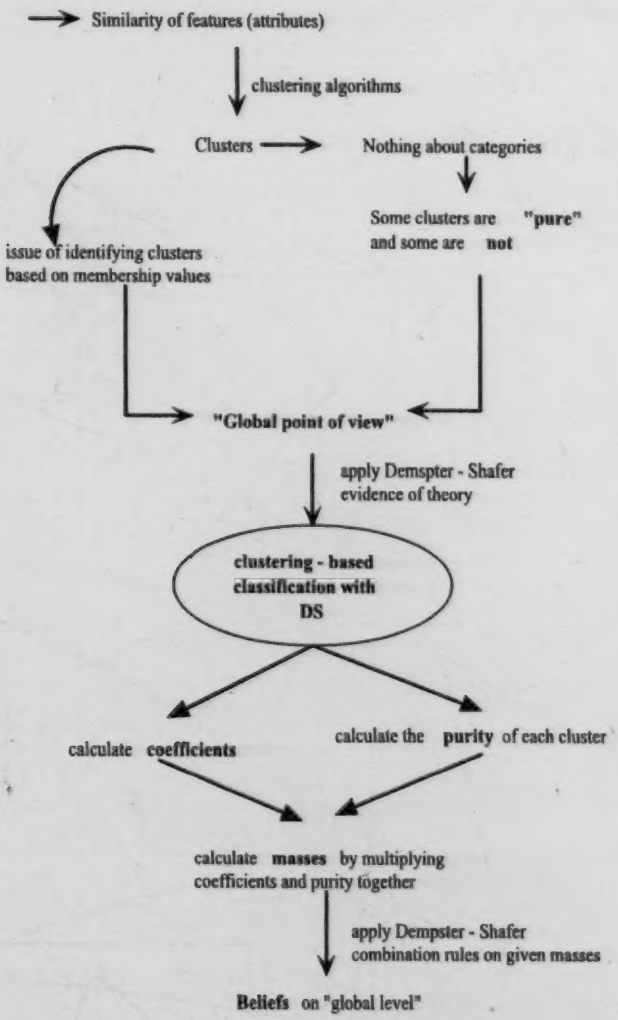

Figure 4.3: Process of obtaining global information

4.4 COMBINING LOCAL AND GLOBAL BELIEFS

As mentioned in the overview part, the work focuses on processing a training set from the global and local points of view, and to build a classifier based on the results of this processing. Therefore, in order to obtain a classification that combine results on both global level and local level, the combination of global beliefs and local beliefs is necessary. In this case, the process of fusing global beliefs and local beliefs is implemented exactly the same as the process described in the global beliefs part – firstly the homogenous beliefs are combined, then conflicting ones. The obtained results are named total beliefs. Total beliefs represents the final conclusion regarding the assignment of a specific category to a new data point.

One of the difficulties of performing clustering is finding a "true" number of clusters. We simply do not know how many clusters of data are in the processed data. The most common method that addresses this problem is based on a "try - and - error" approach. The clustering algorithm is run multiple times, and after each run quality of clusters is estimated using different criteria. For example, the quality can be estimated by calculating two indexes: average compactness of clusters, and distribution of clusters. The "optimal" number of clusters is found by plotting the values of these indexes against a number of clusters, and finding a spot that has the most desirable values of both indexes. There are also more formal approaches that can be applied to find a proper number of clusters that require a number of assumptions and complicated calculations [25]. An interesting comparison of

different methods in presented in [13].

The approach proposed in the thesis does not have the problem of finding a "true" number of clusters. The main concern of the approach is purity of clusters, and this is not directly related to the issue of finding a proper number of homogenous groups of data points. The proposed methods is more interesting in finding groups of data that belong to a single category, and are homogenous at the same time. Of course, this issue is not trivial because we do not know how many of such groups exist in the processing data. In order to mitigate that problem, we propose that clustering is performed a couple of times – each time with a different number of clusters – and the results of each clustering become a part of the solution. This means that there are a couple of sets of global beliefs. Each set of global beliefs is the results of a single clustering run. Again, applying the Dempster - Shafer combination rules, as stated in the first paragraph of this section, to combine local and a number of global beliefs. The obtain beliefs– so–called–the ultimate beliefs, and they are the final results of the proposed method.

CHAPTER 5

EXPERIMENTAL RESULTS

This chapter reports several experimental results to demonstrate the effectiveness of the proposed model. The proposed classifier is simulated on three different data sets where two data sets are synthetic and one is the real data set.

- The first experiment is performed on two dimensional data set to illustrate the key idea of the proposed approach. Visualizations of the data set and statistics corresponding with the final results in each case are depicted and reported.
- The second experiment is simulated on more a complex data set where the size, dimensions and categories of the data set both increase. Statistics of final classification outcomes, as well as classification results achieved from global and local levels are also reported.
- The third experiment is simulated on the real data set which is taken from the UCI data mining repository. Some reports regarding the final classification

results are also given out.

Before presenting the results of some of these experiments, practical issues related to the definitions of name of beliefs obtained in each step of an experiment, interpretations with respect to classification results, as well as the outline organized in each experiment, need to be addressed.

First of all, name of beliefs obtained in each step of a single experiment is defined as follows:

- **Global beliefs** represents for beliefs obtained from global point of view,
- **Local beliefs** denotes for beliefs obtained from local point of view,
- **Total beliefs** represents for beliefs after combining global beliefs and local beliefs together,
- **Ultimate beliefs** represents for beliefs achieved after combining total beliefs from the previous step together, regardless of number of clusters.

Secondly, classification results including statistics about the correct classification, misclassification and undetermined classification are listed in the same table corresponding to each simulation.

- **Correct classification** is represented by percentage which is the division of number of samples which are assigned proper categories over total number of samples in the testing set.

- **Misclassification** is denoted by percentage which is the division of number of samples which are assigned wrong classes over total number of samples in the testing set.

- **Undetermined classification** is also defined by percentage. This percentage is calculated by taking total number of samples where we do not have enough confidence to assign them in one of existing groups over total number of samples in the testing set.

Finally, the outline is organized consistently for all experiments where for every experiment, the first section deals with data description for real data sets and data generation in the case of artificial data sets, the second section provides statistics regarding to results obtained from each step of the proposed classifier.

5.1 Synthetic Data

5.1.1 Data Generation

The following experiment provides general understandings of how the classifier is constructed and how the training and testing sets are applied in each step of the classifier. We considered a three-category problem (with equal prior) and two Gaussian random feature vectors with following characteristics:

$$\mu_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 1.0 \\ 0.5 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 0.20 \\ 0.75 \end{pmatrix}$$

$$\sigma_0 = \begin{pmatrix} 0.25 \\ 0.40 \end{pmatrix}, \quad \sigma_1 = \begin{pmatrix} 0.35 \\ 0.45 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0.45 \\ 0.25 \end{pmatrix}$$

Training and testing sets of 750 and 300 samples respectively are generated independently. Illustrations of training and testing data sets are plotted in Figure 5.1 and Figure 5.2

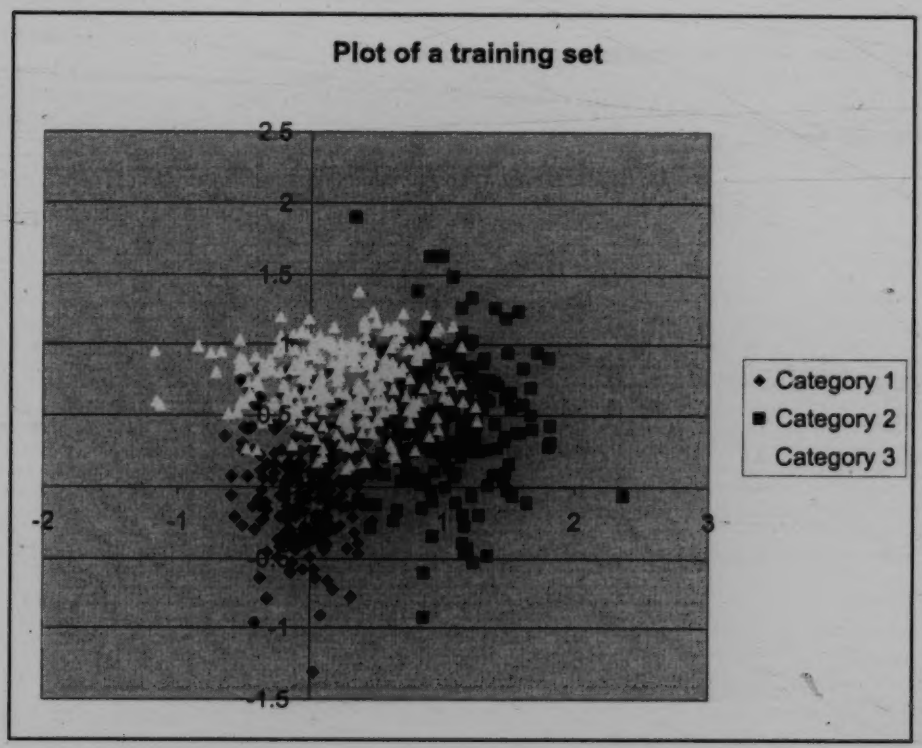

Figure 5.1: Illustration of the training set

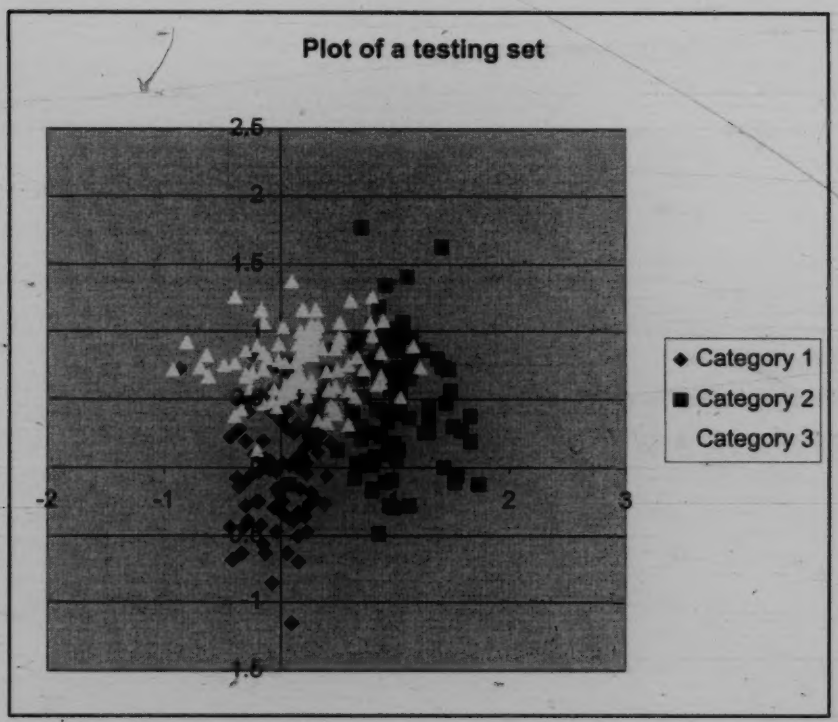

Figure 5.2: Illustration of the testing set

For both training and testing set, the shapes denote for samples belonged to each category as follows: diamond shapes denote for samples belonged to category 1, rectangular shape denote for samples belonged to category 2, and triangular shapes represents for samples belonged to category 3.

According to two figures, we can observe that there are quite number of over-lapped samples among three categories.

5.1.2 Results

As the first step in the process of building the proposed classifier, FCM algorithm is firstly applied to extract information, such as the purity of clusters, also coordinates of cluster centroids, from the training set.

Parameters used for the FCM technique in this experiment are specified as follow: $m = 1.5$, number of clusters = 3, threshold = 0.6. Clusters obtained by the FCM method are shown in the Figure 5.3, where each oval represents approximately for the boundary of each cluster.

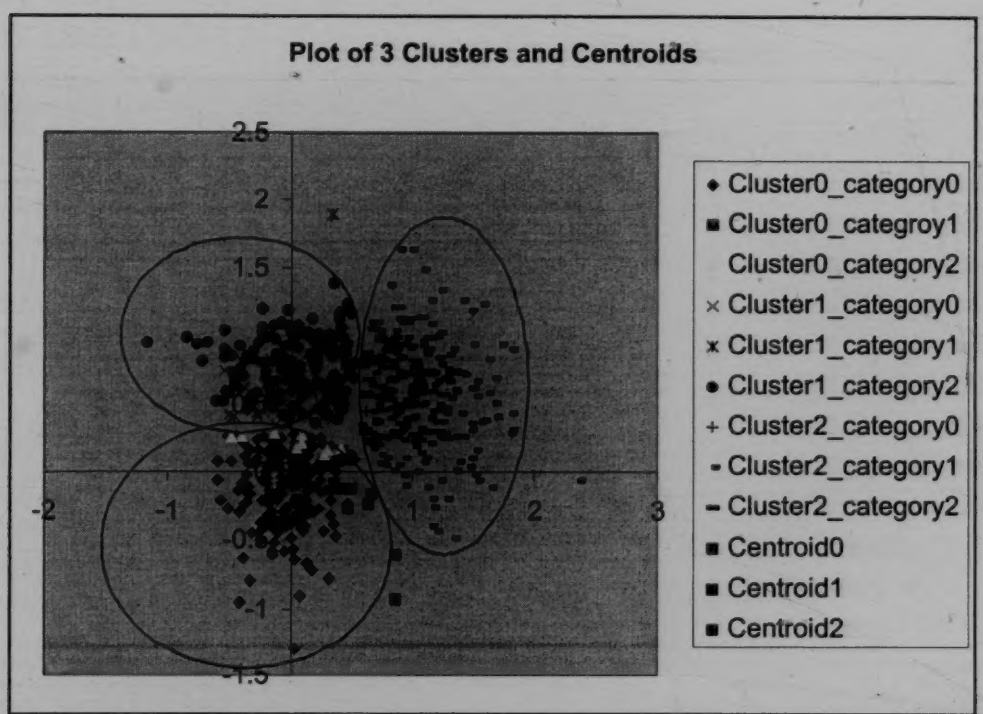

Figure 5.3: Results obtained by the FCM method with 3 clusters

The coordinates of each centroid associated with each cluster are specified as:

centroid 0 = (0.0202, -0.1013),

centroid 1 = (0.0578, 0.7309),

centroid 2 = (1.0123, 0.5752).

A statistic about the purity of each cluster is represented in the table 5.1:

	Category 0	Category 1	Category 2
Cluster 0	0.8911	0.0646	0.0446
Cluster 1	0.2325	0.0395	0.7281
Cluster 2	0.0038	0.8007	0.1954

Table 5.1: Statistics of purity for 3 clusters

The notion of global viewpoint is illustrated clearly with the assistance of two points in the testing set. Two testing points with coordinates = **(-0.27566, 0.431898)** and **(0.367304, 0.238276)**, respectively are taken randomly from the testing set.

Illustrations of global point of view idea can be displayed in Figure 5.4 and Figure 5.5:

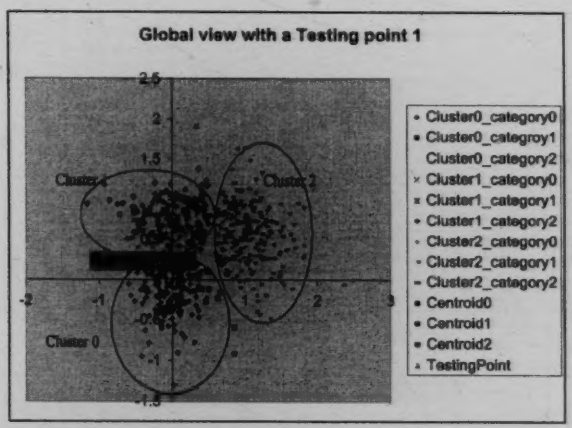

Figure 5.4: The illustration of global level with a testing point 1

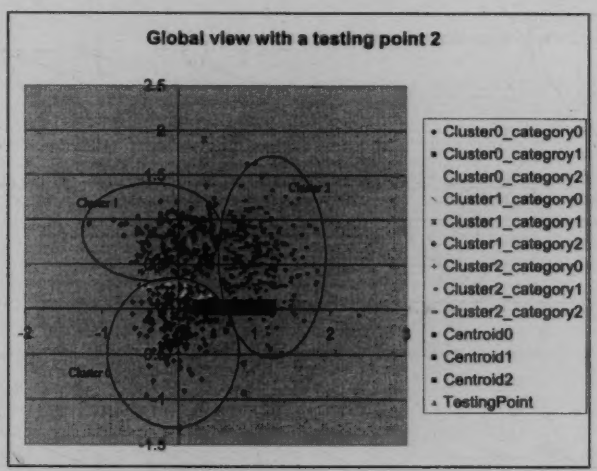

Figure 5.5: The illustration of global level with a testing point 2

By looking at the Figures, we can observe that the testing point 1 should belong to cluster 1 because of its shortest distance from itself to the centroid 1, and the testing point 2 should belong to cluster 0.

The classification task is perfectly done if there is only one class associated

with all data points in a cluster. However, according to the statistic of the purity for each cluster, we do not have enough evidence to assign any label either for the testing point 1 or the testing point 2.

Thus, in order to enhance the level of confidence in assignment tasks for unknown samples on global level, outputs extracted from FCM algorithm is regarded as pieces of evidence which then are merged by means of Dempster–Shafer combination rules.

A statistic related to the final classification results achieved from the global level for the testing set is shown in table 5.2:

Correct Classification	71.67%
Misclassification	28.30%
Undetermined Classification	0.00%

Table 5.2: Statistic of classification results on global level with 3 clusters

In the second step, the idea of local point of view is represented. In this step, the modified k-NN classifier is applied in order to find nearest neighbors within a circle with a given radius. The radius is selected for this experiment is 0.075. Similar to the idea of global point of view, the illustrations of local point of view are plotted corresponding with two specified testing points above in Figure 5.6 and Figure 5.7:

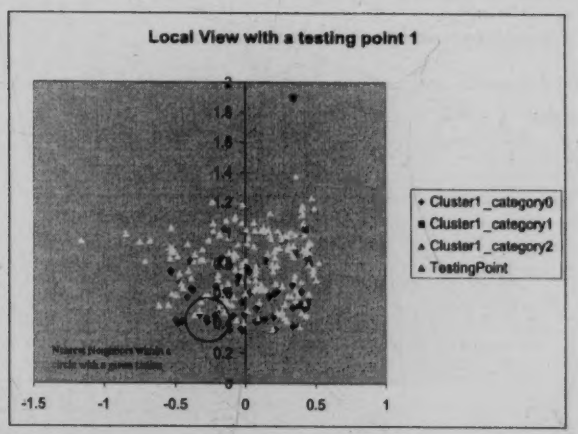

Figure 5.6: The illustration of local view with a testing point 1

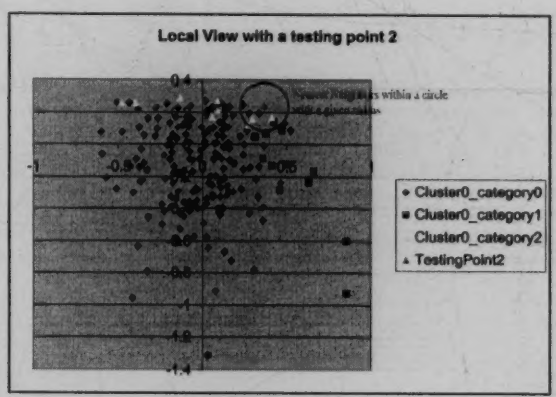

Figure 5.7: The illustration of local view with a testing point 2

As can be seen in two figures, local view reveals more information related to classification for two testing points based on the labels associated with their nearest neighbors. Again, by applying appropriate Dempster–Shafer combination rules on different kinds of masses, local beliefs are achieved.

Numerical facts represented for the classification results corresponding with a testing set on local level are shown in the table 5.3:

Correct Classification	62.33%
Misclassification	18.00%
Undetermined Classification	19.67%

Table 5.3: Statistic of classification results on local level with 3 clusters

The construction of the classifier is totally completed when merging local beliefs and total beliefs together by means of Dempster–Shafer combination rules, resulting in total beliefs. Based on total beliefs, final classification results for samples under consideration in the training set is concluded.

A statistic related to final classification results are reported in the table 5.4:

Correct Classification	65.33%
Misclassification	18.00%
Undetermined Classification	16.67%

Table 5.4: Classification results based on total beliefs with 3 clusters

The work is extended based on the fact that some points in the training set might not be treated well with a low number of clusters, resulting in the low accuracy in classification task. In other words, with higher number of clusters, FCM algorithm could assign some data points in the training set into proper groups. As the result, the labeled assignments for patterns in the testing set can be more accurate. Hence, our research works on that idea by increasing number of clusters to 6 and 9 for the FCM step in the global level.

Figure 5.8 and Figure 5.9 shows how the training set is treated with 6 clusters and 9 clusters, respectively:

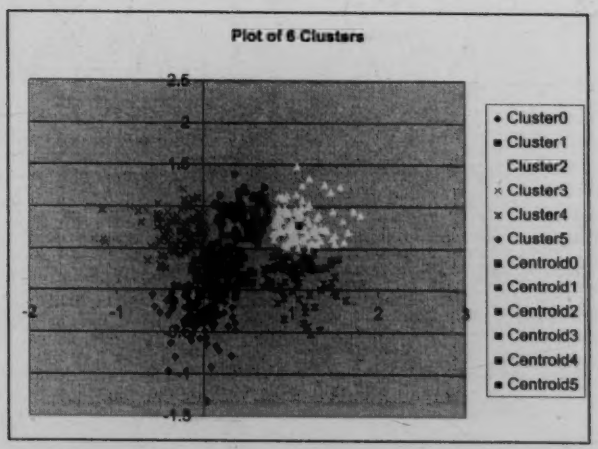

Figure 5.8: Results of 6 clusters obtained by FCM method

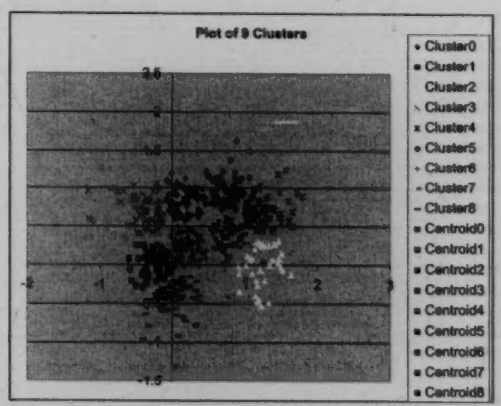

Figure 5.9: Results of 9 clusters obtained by FCM method

The same procedure simulated with 3 clusters is applied to yield the total beliefs in the case of 6 clusters and 9 clusters.

Statistics corresponding with classification results for the testing set based on obtained total beliefs for 6 clusters and 9 clusters are shown respectively in table 5.5 and table 5.6:

Correct Classification	75.00%
Misclassification	5.33%
Undetermined Classification	19.67%

Table 5.5: Classification results based on total beliefs in the case of 6 clusters

Correct Classification	74.00%
Misclassification	24.33%
Undetermined Classification	1.67%

Table 5.6: Classification results based on total beliefs in the case of 9 clusters

Ultimately, the final beliefs from three simulations are combined to observe how better in the final classification of the classifier is.

The ultimate statistics about the labeled assignment for the same testing set is described in the table 5.7:

Correct Classification	75.00%
Misclassification	25.00%
Undetermined classification	0.00%

Table 5.7: Classification results obtained from ultimate beliefs

The classification accuracy of the proposed classifier when combining three different numbers of clusters improves around 10% comparing with using only 3 clusters in the global level.

5.2 Synthetic Data

5.2.1 Data Generation

The intention of the second experiment is to observe how generic of the model is when coping with a more complex data set. The proposed approach faces with a six-category problem (with equal prior) and four Gaussian random feature vectors with following characteristics:

$$\mu_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 1.05 \\ 0.65 \\ 0.0 \\ 1.35 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} -0.20 \\ -0.85 \\ 1.20 \\ -0.95 \end{pmatrix},$$

$$\mu_3 = \begin{pmatrix} -0.50 \\ 0.75 \\ -0.25 \\ 0.60 \end{pmatrix}, \quad \mu_4 = \begin{pmatrix} 0.00 \\ 0.75 \\ -0.20 \\ -1.50 \end{pmatrix}, \quad \mu_5 = \begin{pmatrix} -1.45 \\ -0.50 \\ 1.40 \\ 0.35 \end{pmatrix}.$$

$$\sigma_0 = \begin{pmatrix} 0.50 \\ 0.75 \\ 0.25 \\ 0.40 \end{pmatrix}, \quad \sigma_1 = \begin{pmatrix} 0.30 \\ 0.55 \\ 0.60 \\ 0.40 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0.75 \\ 0.65 \\ 0.45 \\ 0.85 \end{pmatrix},$$

$$\sigma_3 = \begin{pmatrix} 0.35 \\ 0.50 \\ 0.65 \\ 0.40 \end{pmatrix}, \quad \sigma_4 = \begin{pmatrix} 0.35 \\ 0.55 \\ 0.60 \\ 0.55 \end{pmatrix}, \quad \sigma_5 = \begin{pmatrix} 0.65 \\ 0.55 \\ 0.70 \\ 0.55 \end{pmatrix}.$$

Training and testing sets of 1800 and 720 samples respectively are generated independently. Because this data set is a four dimensional data set, therefore, we can not see the distribution of all data samples. However, due to the standard deviations, we can conclude that the data set samples are highly overlapped. Hence, the task of assigning an accurate category for each sample in a testing set is not a trivial task.

5.2.2 Results

Parameters selected to train the new classifier are specified as following:

- A number of clusters selected for the FCM technique in the global level are 6, 9 and 12 clusters.
- The fuzziness degree is selected equal to 1.5 or $m = 1.5$.
- The threshold ultilized to decide the boundary of each cluster is selected equal to 0.6.
- The radius for k-NN classification method in the local level is chosen as 0.4.

Experiments in this part deal with a high dimensional data set, hence, the purity of each cluster, as well as their corresponding centroids are reported to enhance understanding about the distribution of the training set.

The first experiment is simulated with 6 clusters applied on the global level. The purity of each cluster is reported in the table 5.8:

	Category 0	Category 1	Category 2	Category 3	Category 4	Category 5
Cluster 0	0.00	0.00	0.0731	0.00	0.00	0.9269
Cluster 1	0.00	0.00	0.9827	0.00	0.0058	0.0116
Cluster 2	0.1562	0.0033	0.00	0.7980	0.0166	0.0199
Cluster 3	0.0078	0.00	0.00	0.00	0.9922	0.00
Cluster 4	0.8333	0.00	0.0857	0.0429	0.0143	0.0238
Cluster 5	0.0071	0.9929	0.00	0.00	0.00	0.00

Table 5.8: Statistics of purity for 6 clusters

Centroids corresponding with each cluster have coordinates as follows:

centroid 0 = (-1.4558, -0.5382, 1.5411, 0.2775),

centroid 1 = (-0.1638, -0.9062, 1.2022, -1.2616),

centroid 2 = (-0.4455, 0.7473, -0.1662, 0.4961),

centroid 3 = (0.0024, 0.7883, -0.2807, -1.5949),

centroid 4 = (0.0667, -0.3117, 0.1689, 0.0202),

centroid 5 = (1.0015, 0.6849, -0.0036, 1.3642).

The classification results for global level are shown in the table 5.9:

Correct Classification	59.73%
Misclassification	40.27%
Undetermined classification	0.00%

Table 5.9: Classification results on global level with 6 clusters

The next step is to look at the statistic of classification results based on obtained local beliefs from local level.

The classification statistic is displayed in the table 5.10:

Correct Classification	56.53%
Misclassification	8.20%
Undetermined classification	35.27%

Table 5.10: Classification results on local level with 6 clusters

The classification results based on total beliefs is indicated in the table 5.11:

Correct Classification	55.97%
Misclassification	8.20%
Undetermined classification	35.83%

Table 5.11: Classification results of the model with 6 clusters

Similar to the former part, the same procedure is applied to train the model with 9 clusters on global level.

The purity of each cluster is reported in the table 5.12:

	Category 0	Category 1	Category 2	Category 3	Category 4	Category 5
Cluster 0	0.00	0.00	0.233333	0.0111	0.00	0.7556
Cluster 1	0.8739	0.00	0.0910	0.0090	0.0090	0.0100
Cluster 2	0.00	0.00	0.0278	0.00	0.00	0.9722
Cluster 3	0.00	0.00	0.9916	0.00	0.0084	0.00
Cluster 4	0.2040	0.00	0.00	0.7551	0.0102	0.0306
Cluster 5	0.0051	0.00	0.00	0.00	0.9949	0.00
Cluster 6	0.0310	0.0078	0.00	0.9612	0.00	0.00
Cluster 7	0.7368	0.00	0.00	0.0395	0.2237	0.00
Cluster 8	0.00	1	0.00	0.00	0.00	0.00

Table 5.12: Statistics of purity corresponding with 9 clusters

Centroids corresponding with 9 clusters has coordinates as follows:

centroid 0 = (-0.8981, -0.6983, 1.5424, -0.0091),

centroid 1 = (0.2046, -0.5887, 0.1778, 0.0019),

centroid 2 = (-1.8117, -0.4592, 1.4813, 0.3783),

centroid 3 = (-0.0991, -0.9521, 1.1797, -1.4489),

centroid 4 = (-0.4509, 0.4027, 0.2763, 0.5249),

centroid 5 = (0.0014, 0.8108, -0.3138, -1.7234),

centroid 6 = (-0.4655, 0.9059, -0.4953, 0.5884),

centroid 7 = (0.1019, 0.4819, -0.0283, -0.3338),

centroid 8 = (1.0379, 0.6877, -0.01163, 1.4119).

Taking a close look at the purity table in the case of 9 clusters, even though each cluster is still not pure, each cluster contains a dominant category inside itself. This result affects substantially to the outcomes on global level. It promises that the accuracy of classification on global level will be elevated comparing with the case of 6 clusters.

The table 5.13 below represents for the statistic regarding to the assignment results for the testing set on global level:

Correct Classification	62.63%
Misclassification	37.36%
Undetermined classification	0.00%

Table 5.13: Classification results on global level with 9 clusters

A statistic of the classification accuracy on local level is displayed in the table 5.14:

Correct Classification	56.53%
Misclassification	8.20%
Undetermined classification	35.27%

Table 5.14: Classification results on local level with 9 clusters

A statistic of the classification accuracy when combining local beliefs and global beliefs together is reported in the table 5.15.

Correct Classification	57.36%
Misclassification	8.47%
Undetermined classification	34.17%

Table 5.15: Classification results of the model with 9 clusters

Three statistic tables report all the results regarding to the classification results in the case of applying 12 clusters to global level of the model. Both three tables show the statistic regarding to the accuracy of assigning proper labels for the training set on global level, local level and total level, respectively.

Correct Classification	54.44%
Misclassification	45.55%
Undetermined classification	0.00%

Table 5.16: Classification results on global level with 12 clusters

Correct Classification	56.53%
Misclassification	8.20%
Undetermined classification	35.27%

Table 5.17: Classification results on local level with 12 clusters

Correct Classification	69.44%
Misclassification	9.72%
Undetermined classification	20.83%

Table 5.18: Classification results of the model with 12 clusters

The final table 5.19 for the second experiment demonstrates the statistic of the classification accuracy after combining final outcomes obtained from the model with 6, 9 and 12 clusters.

Correct Classification	76.80%
Misclassification	11.52%
Undetermined classification	11.66%

Table 5.19: Classification results of the model based on ultimate beliefs

As a final result, the classification accuracy enhances around 20.83% comparing with the final classification in the case of 6 clusters, 19.44% in the case of 9 clusters and 7.44% with the case of 12 clusters applied on the global level.

5.3 Real World Data

5.3.1 Data Description

In this section, the model will be trained with a real data set which was taken from the UCI data mining resources. UCI is a repository of databases, domain theories and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithm [27]. The repository presented the following characteristics: Database Name, Number of Instances (i.e examples, data points, observations); Name of Features (i.e. dimensions, attributes); Num-

ber of classes (assuming a discrete class variable); Percent of features that have continuous/integer values, Percent of features that have nominal values; Missing features (Yes/No); Highest Reported Accuracy(taken from "Past Usage"); Percent of Instances in the Majority Class [15].

The selected data set concerns an image segmentation problem. The data was created by Vision Group at University of Massachusetts in November, 1990. The samples were drawn randomly from a database of 7 outdoor images. Each instance is a 3×3 region. There are 210 samples for the training set and 2100 samples for the testing set. Each sample corresponds with 19 attributes and no missing attribute is found. There are 7 categories which are names as brickface, sky, foliage, cement, window, path and grass.

The aim is to build a classifier based on the data set to predict classes for samples in the testing set, as well as to gain the level of confidence when assigning a category for a specific testing sample. For the purpose of simplification, we convert character–type classes to number–type classes. Hence, the data set still contains 7 categories but with the class distribution from 0 to 6 where each number corresponds with each name mentioned previously.

5.3.2 Results

After the first two experiments, we already have ideas how the classifier is constructed, how the data set is exploited in the model and where outcomes are

obtained. In this part, the identical process with identical parameters except the radius in local level is applied into the training set and testing set in order to build the model and to obtain the final outcomes. The radius for this particular problem is selected equal to 35.00. Hence, there are three experiments corresponding with three different number of clusters.

Three tables, in particular, table 5.20, table 5.21, table 5.22, below are statistics about the classification results which are calculated based on outcomes obtained from each step in each experiment, respectively.

	Global level	Local Level	Total Level
Correct Classification	14.29%	56.57%	62.57%
Misclassification	85.71%	14.57%	36.71%
Undetermined classification	0.00%	28.85%	0.72%

Table 5.20: Final classification results of the model with 6 clusters

	Global level	Local Level	Total Level
Correct Classification	29.33%	56.57%	65.95%
Misclassification	70.67%	14.57%	26.23%
Undetermined classification	0.00%	28.85%	7.80%

Table 5.21: Final classification results of the model with 9 clusters

	Global level	Local Level	Total Level
Correct Classification	39.67%	56.57%	65.67%
Misclassification	60.33%	14.57%	22.19%
Undetermined classification	0.00%	28.90%	12.14%

Table 5.22: Final classification results of the model with 12 clusters

The final statistic regarding to classification results is shown in the table 5.23,

where ultimate beliefs are obtained by combining total beliefs from previous simulations:

Correct Classification	68.05%
Misclassification	31.95%
Undetermined classification	0.00%

Table 5.23: Final classification results of the model based on ultimate beliefs

According to the statistics about the classification results from three experiments, we can strongly prove that the proposed classifier gives out a better classification when combining multiple number of clusters together. Moreover, the model shows that not only it assigns proper categories for unknown samples but also provides the degree of confidence for samples which are not misclassification, just we do not have enough certainty to classify them into any groups.

CHAPTER 6

CONCLUSIONS

The amount of information and data stored in standalone databases, and on the web is increasing exponentially. In order to make this data useful there is a need for different data processing methods and techniques. One group of such techniques focuses on constructing data models that can be used for classification purposes. This thesis proposes a novel approach for building classifiers.

The proposed approach introduces the concepts of local and global analysis of data. To "look" at data locally the kNN method is used – in this case any knowledge about a new data point is inferred based at a local level where only closest neighbors are taken into consideration. In the case of a "global" point of view – a clustering is applied. Construction of clusters allows us to identify an "identity" of a new data point based on its location in reference to groups of homogenous data. Purity of these groups is also taken into consideration. Both levels of analysis are "glued"

using elements of Dempster - Shafer evidence theory. KNN and clustering have been modified in order to accommodate concepts of belief masses that are required in order to use Dempster - Shafer combination rules that provide us with the final classification result.

The application of evidence theory has introduced an interesting concept to the classification process: besides original categories that are given with the data, a new category – undetermined – is added. When a new data point belongs to this category it is indicated that there is not enough evidence supporting a statement that of a new data point belongs to one of the original categories.

The proposed method has been applied for building data classifiers for a number of data sets. The results obtained from these experiments indicate that the method is very promising.

In overall, the presented method contributes to the process of building classification systems. Each of the components of the method brings its own flavor to the classification process: a global view is provided by clustering, and a local view is covered by kNN technique. Both methods complement each other. One of the valuable aspects of applying two different techniques for data analysis is to increase users confidence in obtained results. This happens when both techniques result in the same classification. Of course conflicting classifications also occur, in such situation user is provided with an indication that the system is not capable of providing

a reliable classificaiton.

Several areas of possible future research are presented below:

- different methods can be consider for evaluation of belief masses:
  - for kNN method: the concept of circle can be revisited – our early investigations have indicated that it would be valuable to apply some techniques leading to automatic adjustment of circle's diameter; different algorithms for calculating values of individual belief masses should be explored;
  - for clustering based classification: as in the case of kNN, different approaches used for calculating belief masses should be investigated;
- influence of threshold value used for identifying cluster borders should be investigated;
- different clustering techniques, for example, hierarchical one, should be tried;
- the issue of number of clusters should be analyzed further;
- the presented results focused only on accuracy of classification, other measures such as precision and recall should be also investigated, it would be interesting to see if it is possible to adjust the combination of individual classifications to emphasis one of them;
- results obtained by kNN and clustering based classification are treated equally (we can say that they are combined with the same weights), it would be very

interesting to introduce some weights that would control "contribution" of each method to the final classification.

REFERENCES

[1] Ahmadzadeh, M.R., and Petrou, M., Use of Dempster-Shafer theory to combine classifiers which use different class boundaries, Pattern Analysis and Applications 6 (2003) 41-46.

[2] Al-Ani, A., and Deriche, M., A New Technique for Combining Multiple Classifiers using the Dempster-Shafer Theory of Evidence, Journal of Artificial Intelligence Research 17 (2002) 333-361.

[3] Babcock, C., Parallel processing mines retail data, Computer World 6 (1994).

[4] Bezdek, J., Pattern Recognition with Fuzzy Objective Function Algorithms, Plenun Press, New York (1981).

[5] Berkhin, P., Survey of Clustering Data Techniques, http://www.ee.ucr.edu/ barth/EE242/clustering–survery.pdf

[6] Cover, T., and Hart, P., Nearest Neighbor pattern classification, IEEE Trans. On Information Theory, 13 (1967), 21-27.

[7] Dasarathy, B.V., Nearest-Neighbor Classification Techniques, IEEE Computer Society Press, Los Alomitos, CA (1991).

[8] Denoeux, T., A k-nearest neighbor classification rule based on Dempster-Shafer theory, IEEE Transactions on Systems, Man and Cybernetics 25(5) (1995) 804-813.

[9] Denoeux, T., Analysis of evidence-theoretic decision rules for pattern classification, Pattern Recognition 30(7) (1997) 1095-1107.

[10] Denoeux, T., A neural network Classifier Based on Dempster-Shafer Theory, IEEE Trans. On Systems, Man, and Cybernetics Part A 30(2) (2000) 131-150.

[11] Gaber, M.M., et al, A Survey of Classification Methods in Data Streams, Springer US(2007).

[12] Hashemi, Ray R., and Bahar, H., Decoupling of Clustering and Classification Steps in a Cluster–Based Classification, Proceedings of the Fourth International Conference on Machine Learning and Applications, IEEE (2005).

[13] Hardy, A., On the number of clusters, Computational Statistics and Data Analysis, 23(1) (1996), 83-96.

[14] Heywood, M., http://users.cs.dal.ca/ mheywood/CSCI6506/HandOuts/N03-GA-Cmeans.pdf.

[15] Igbide, E., Knowledge extraction on software engineering data, MSc thesis (2003), Department of Electrical and Computer Engineering, University of Alberta.

[16] Jain, A.K., Murty, M.N., and Flynn, P.J., Data Clustering: A review, ACM Computing Surveys, September (1999), 31(3).

[17] Mandler, E. and Schurmann, J., Combining the classification results of independent classifiers based on the Dempster-Shafer theory of evidence, in Gelsema, E. and Kanal, L. (eds), Pattern Recognition and Artificial Intelligence (1988) 381-393.

[18] Mirkin, B., Clustering for Data Mining: A Data Recovery Approach, Taylor and Francis Group, LLC(2005).

[19] Nascimento, S., Mirkin, B., and Moura–Pires, F., A Fuzzy Clustering Model of Data and Fuzzy c–Means, The 9th IEEE Conference on Fuzzy System (FUZZ-IEEE 2000),(to appear).

[20] Parsons, S., and Hunter, A., A Review of Uncertainty Handling Formalism, Applications of Uncertainty Formalisms(1998), 8-37.

[21] Pedrycz, W., and Gomine, F., An Introduction to Fuzzy Sets, The MIT Press (1998).

[22] Rogova, G., Combining the results of several neural network classifiers, Neural Networks 7 (1994) 777-781.

[23] Shafer, G., A Mathematical Theory of Evidence, Princeton University Press (1976).

[24] Sentz, K., and Ferson, S., Combination of Evidence in Dempster–Shafer Theory, Sandia Report, SAND2002–0835, Unlimited Release, Printed April(2002).

[25] Sugar, C.A., and James, G.M., Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach, Journal of the American Statistical Association, 98(463) (2003), 750-763.

[26] Turcan, A., Ocelikova, E., and Madaraz, L., Fuzzy C–means algorithms in remote sensing, http://www.bmf.hu/conferences/SAMI2003/Ocelikova.pdf.

[27] UCI Machine Learning Repository, http://mlearn.ics.uci.edu/MLRepository.html.

[28] Way, J., and Smith, E.A., The Evolution of synthetic aperture radar systems and their progression to the EOS SAR, IEEE Transactions on Geoscience and Remote Sensing 29 (1991) 962–985.

[29] Wang, W., Predictive modeling based on classification and pattern matching methods, MSc thesis, Simon Fraser University, May(1999).

[30] Weiss, S.M., and Kulikowski, C.A., Computer Systems that Learn: Classification and Prediction Methods from Statistic, Neural Nets, Machine Learning, and Expert Systems. Morgan Kaufman(1991).

[31] Li, Y., Dai, H., Reducing Uncertainties in Data Mining, apsec, p.97, Fourth Asia–Pacific Software Engineering and International Computer Science Conference (APSEC/ICSC) (1997).

[32] Zadeh, L.A., A Fuzzy–Algorithmic Approach to the Definition of Complex or Imprecise Concepts, Int.J.Man–Machine Studies, 8 (1976), 249–291. Reprinted with permission. Copyright 1976 by Academic Press Inc (London) Ltd.

[33] Zadeh, L.A., Fuzzy sets, Information and Control, 8(1965), 338- 353, .

[34] Zadeh, L.A., A fuzzy set–theoretic Interpretation of Linguistic hedges, Journal of Cybernetics, 2:2 (1972), 4–34, Reprinted with permission. Copyright 1972 by Hemisphere Publishing Corporation, a subsidiary of Harper & Row, Publishers, Inc.

[35] Zaiane, O., http://www.cs.ualberta.ca/%7Ezaiane/courses/cmput690/slides/Chapter8/sld001.htm.

[36] Zeng, et al., CBC: Clustering Based Text Classification Requiring Minimal Labeled Data, ICDM (2003), 443-450.